The representation of mood in primate anterior insular cortex

Nicole C. Rust[1], You-Ping Yang[2], Catrina M. Hacker[1], Veit Stuphorn[3]
[1]Department of Psychology, University of Pennsylvania, Philadelphia, USA
[2]School of Medicine, National Defense Medical Center, Taipei, Taiwan
[3]Department of Neuroscience, Johns Hopkins University, Baltimore, USA

Significance Statement:

How does the brain shape mood? Answering this question is difficult because measures of mood require asking humans how they feel, but our best measures of brain activity are restricted to nonhuman animals. To overcome this challenge, we developed a model to link across these different types of data as human happiness fluctuated following wins and losses during a gambling game, and monkeys played the same game. We found a strong neural correlate of human happiness reflected in the persistent responses of neurons in monkey anterior insular cortex, a brain area implicated in the conscious experience of feelings. These results suggest that the brain may convert experiences into mood by integrating experiences through a network of recurrently connected neurons.

Abstract:

Understanding how the brain reflects and shapes mood requires resolving the disconnect between behavioral measures of mood that can only be made in humans (typically based on subjective reports of happiness) and detailed measures of brain activity only available in animals. To achieve this, we developed a universal mood model to predict behavioral fluctuations in human subjective happiness as individuals experienced wins and losses during a gambling task. Next, we investigated how this operationalization of mood was reflected in the brains of two monkeys engaged in the same gambling task. We found a remarkable alignment between human mood model signatures and the persistent responses of units in monkey anterior insular cortex, including a matched timescale of integration across events. In comparison, the same signatures were only weakly reflected in dorsolateral prefrontal cortex, suggesting that insular mood representations do not trivially follow from a signal (like arousal) broadcast to all higher brain areas. These results are consistent with a model in which the brain transforms experiences into mood by integrating events through a recurrently connected network of excitatory and inhibitory pools of neurons. These are among the first detailed insights into the nature of putative mood representations in the primate brain.

Introduction:

Of all the functions of the brain and mind, mood is among the most mysterious. In comparison with acute emotions that are often targeted at something specific (like the fear of a tiger), moods are typically more diffuse and slowly changing, as mood reflects an integration across multiple events (1–5). Understanding mood is important, given the vast unmet need of individuals with mood disorders including major depression and bipolar disorder. A remarkable 21% of adults

will experience a mood disorder at some point in their lives (6), and a third of those will be resistant to existing treatments (7). The development of new and improved behavior-based and brain-based interventions for mood disorders could benefit from a better understanding of how mood is supported and shaped by the brain (8, 9).

To date, researchers have learned a considerable amount about how mood is influenced by what happens at the highly detailed molecular level (e.g. by serotonin neurotransmission, (10)). They've also made progress toward understanding the broader scale brain networks involved in mood via fMRI (e.g., 10). In comparison, remarkably few studies have focused on how patterns of brain activity exhibited by populations of individual neurons drive the experience of mood, how processing in the brain converts external and internal experiences into those brain activity patterns, and how these mood representations influence behavior. To fill this gap, we will need to overcome impediments that make mood more challenging to study than other brain functions such as perception, memory, and decision making.

Many of the challenges associated with understanding mood follow from the fact that mood is a subjective experience for which there is no external ground truth — no one knows how happy an individual is but themselves. Consequently, mood cannot be studied by benchmarking against performance on objectively defined tasks in the same way many other brain functions are studied. Likewise, mood is difficult to study in animals because we lack good ways to measure how they feel (2). While behavioral tests such as the forced swim test have been used to study mood and depression in animals, their relevance to understanding mood in humans is controversial (12). Similarly, inferring mood from innate behaviors such as facial expressions is thought to be problematic (13, 14).

At the same time, the types of densely sampled, high spatial and temporal resolution measures of neural activity that are best suited to investigating neural representations are difficult to obtain in humans. While meaningful data can be gathered from human patients with electrodes implanted for epilepsy or other clinical purposes, those measurements are often sparse and are typically lower-resolution measures of activity that may not reflect the action potentials that the brain uses to communicate within and between brain areas (15). Similarly, while a handful of studies have made notable progress in "decoding mood" from human neural activity (16–18), all have taken a clinically inspired engineering approach focused on the accuracy with which mood can be decoded from the human brain as opposed to focusing on the nature of the neural code for mood itself.

For these reasons, we know little about how answers to questions like, "How happy are you right now?" are reflected and shaped by processing in the brain, analogous to our deeper understanding of other brain functions. Here, we sought to overcome these challenges by applying an approach that seeks to understand mood by using models to link across data of different types.

Results:

To study mood and its neural correlates in a rigorous way, we need to operationalize its definition, measure it, and systematically modulate it. Our operationalization and modulation of

mood captures three generally well-agreed-upon characteristics of it (1): 1) measures of mood most often probe subjective reports of happiness; 2) mood reflects an integration of events as opposed to individual experiences (1); and 3) mood is modulated bidirectionally by both positive and negative events. Below, we refer to the operationalization we use here as "mood" for ease of reading (as opposed to the implication that mood is a singular construct, driven by a single process).

Our approach builds on the work of Rutledge et al. (19), who introduced a way to model and thus predict how an individual's mood will fluctuate in the context of a gambling task. As an overview, they modulated mood via monetary wins and losses on gambling trials, and they measured mood with subjective happiness probes inserted every few trials. They then used this behavioral data to fit a model that predicted mood based on the wins and losses each individual experienced, and that model provided reasonable accounts of subjective happiness fluctuations.

Characteristic of mood, Rutledge et al. (19) found that individuals' answers to the question, "How happy are you right now?" depended not just on events on the last trial but an integration of events across ~7 trials, on average across subjects. They also determined via fMRI that this operationalization of mood was reflected in anterior insular cortex (AIC). In comparison, fMRI activation of the striatum was correlated with gambling trial outcomes but not while answering questions about happiness itself. These results are consistent with multiple lines of evidence that tie AIC to the conscious experience of affective feelings, including mood (20). However, they do not answer the question of how happiness is reflected in this brain region at the level of individual AIC neurons, nor do they shed insight into how mood is computed by the brain.

For answers, we analyzed one of the few existing datasets of AIC activity in the nonhuman primate brain, measured at single-unit resolution. In this study (21), rhesus monkeys performed a gambling task similar to that investigated by Rutledge et al. (19). As an overview of our report, we begin by developing a class of mood models to predict human behavioral mood modulation. Next, we describe the neural correlates of this universal mood model in the monkey brain, focusing on AIC following the fMRI evidence cited above. Here, we assume that the representation of human subjective happiness can be meaningfully studied by probing neural representations in monkeys, contingent on finding straightforward and robust neural correlates.

The drawback of our approach is that it could, in principle, fail — either because happiness fluctuations are dominated by individual differences or because the monkey brain does not compute it. The benefit of this approach is that if we find a nontrivial alignment between human behavioral estimates of happiness and monkey neural activity, it could provide a putative neural correlate of mood that we can build upon to meaningfully understand how the brain shapes mood. With this approach, we found a remarkable alignment between the mood behavioral signatures derived from human subjective accounts and the responses of individual AIC units during the behaviorally unconstrained period between trials. These mood representations were reflected in the activity of a large fraction of the AIC population. In comparison, the same mood signatures were nearly absent in dorsolateral prefrontal cortex.

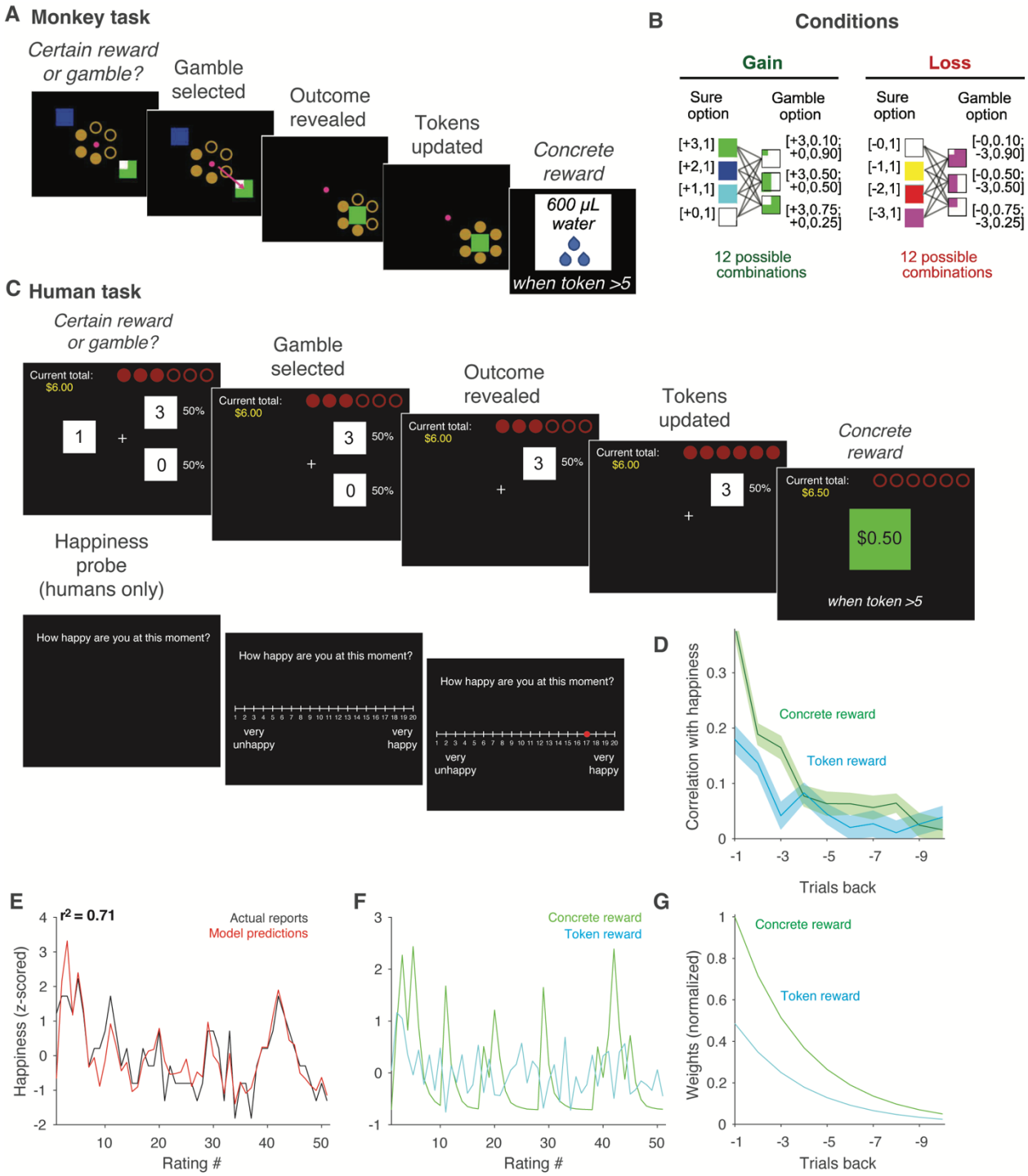


Figure 1. Behavioral paradigm and mood model predictions. A) On each trial of the monkeys' gambling task, monkeys selected between a certain reward (upper left) and a gamble (lower right) for a token reward where tokens were reflected as colored blocks that the monkeys had learned to map to their token values. On each trial, -3 to 3 tokens were rewarded. When the token number reached 6, the monkeys received a juice reward. B) Experimental conditions and their symbols. Trial types were parsed into token wins (Gain) and token losses (Loss) trials, with all possible combinations of 4 certain rewards and 3 gambles for each type. C) The human version of the task was designed to match the monkeys', with the exception that token numbers rather than symbols reflected the options, and humans won money as opposed to juice. A happiness probe was interleaved every 2-3 trials. D) Correlation

between human happiness reports with both monetary rewards and total rewards, demonstrating an influence of both on happiness reaching back ~6-8 trials. Lines indicate means and the shaded region indicates standard error across 44 subjects. **E)** The happiness reports of one human subject (black), along with the universal happiness model prediction of how their happiness would fluctuate (red). **F)** The model predictions in panel E (red), broken down into fluctuations in monetary reward (green) and token number (cyan), which are summed to produce the model predictions in panel E. **G)** The fluctuations in panel F were computed by taking the monetary and token reward events that happened during the course of the experiment, and integrating them through the weights shown. Compare with panel D.

Human behavioral mood models:

Our method for determining the neural correlates of mood fluctuations combined the approaches developed by Rutledge et al. (19) with the details of the gambling task performed by the monkeys during AIC recording, originally reported by (21). In the latter study, the monkeys gambled for wins and losses of 0-3 tokens per trial, and they received a liquid reward when the number of tokens reached 6 (Fig 1A). Specifically, on each trial the monkeys selected between a certain outcome and a gamble between two options weighted by different probabilities, resulting in a change in token number that ranged from -3 (lost) to 3 (gained) on each trial. On some trials, the best choice was obvious, such as a +3 token certain reward versus a +3/0 token gamble with 0.1/0.9 probabilities (where selecting the certain reward was the obvious best choice). On other trials, the best choice was less obvious, such as -1 tokens for certain versus a 0/-3 gamble with 0.5/0.5 probabilities. When the total number of tokens reached 6, the monkey received a juice reward and 6 tokens were subtracted from the token total.

Trials came in two types, parsed by token wins (gains) and losses; altogether, there were 12 types of gain trials and 12 types of loss trials (Fig 1B). In addition, each of the 13 options (both certain rewards and gambles) was presented independently on "forced choice" trials in which the monkey was required to select that option. In total, 37 unique trials were randomly shuffled and each presented once before re-randomization.

The human version of the task was, by design, intended to match the monkeys', with the exception that humans received a $0.50 reward each time the number of tokens reached 6 instead of juice (Fig 1C). For ease, we refer to both liquid (monkeys) and money (humans) as "concrete" rewards to differentiate them from the more abstract "token" rewards that were accumulated to receive concrete ones. In addition, a mood probe was introduced in the human version of the task every 2-3 trials where they were asked to rate their momentary happiness on a scale of 1-20 (Fig 1C). For a minority of individuals, subjective happiness reports were insensitive to the events that happened during the gambling mood induction paradigm (Supp Fig 1). We thus applied a screen to eliminate subjects whose happiness reports were not significantly correlated with concrete rewards that happened on the previous trial, with a generous p-value criterion ($p<0.15$); 11 (of 55) subjects were eliminated and 44 were retained for further analysis. The results we present in this report are qualitatively the same when all 55 subjects were included in these analyses (not shown).

Here we begin by demonstrating the relationship between human subjective happiness reports and the two types of rewards (concrete and token), not just received on the last trial but across a handful of trials. Consistent with the notion that mood reflects an integration of events (as well

as the previous results by Rutledge et al. (19)), reports of subjective happiness were positively correlated with concrete rewards across ~8 trials leading up to a happiness probe on average, across 44 human subjects (Fig 1D, green). Happiness was similarly correlated with token reward across the last ~6 trials (Fig 1D, cyan).

Next, we evaluated several different mood models for their ability to predict happiness fluctuations. To focus on what modulates reward in the gambling task (as opposed to the external factors that set its baseline), happiness reports were z-scored for each subject. Consistent with the intuitions depicted in Fig 1D, one conceptually straightforward model computed subjective reports of happiness as a weighted sum of concrete and token rewards integrated with a fading average over the last ~7-8 trials. We'll refer to this as the "concrete reward and token reward (CRTR) model". Shown in Fig 1E are the happiness fluctuations for one human subject, along with the CRTR model fit to data from 43 human subjects to predict this one individual's held-out data ($r^2 = 0.71$). For this subject, the universal model provided good predictions of how this person's happiness would fluctuate across the gambling task based on the wins and losses they experienced. CRTR model predictions were created as a weighted sum of terms capturing concrete and token reward events (Fig 1F), with the relative weightings over time shown in Fig 1G. Simpler models used to estimate happiness did not perform as well, including ones based on integrated concrete reward alone ($r^2 = 0.58$), integrated token reward alone ($r^2 = 0.24$), or combined concrete and token reward but without integration (i.e. just the events on the previous trial, $r^2 = 0.43$; Supp Fig 1).

This example subject reflected trends in the overall data. On average across all subjects, the CRTR model provided a reasonable account of the main features reflected in the behavioral data (mean $r^2 = 0.38$; Supp Fig 2). Models with fewer terms (such as concrete or token rewards alone) produced worse happiness predictions than the CRTR model (Supp Info, Table 2). Conversely, a subset of models that parsed token-related events into additional terms (such as the tokens won when the certain reward versus the gambling option was selected) provided better predictions (Supp Info, Table 2).

Given the similarity of model fits, coupled with the similarities in the parameterization of these models, we hesitate to declare that there is a singular "best model" that accounts for subjective happiness reports in this data. Rather, a more parsimonious description is that a class of models in which happiness depends on a weighted combination of concrete rewards and token-related events is a decent description of the data overall. We thus focus our analysis of the neural data on the CRTR model due to the conceptual simplicity of that model (Fig 2). We then follow with an exploration of how those results generalize to other, more complex models (Fig 3).

Insight into the neural mechanisms that shape mood

To recap, the CRTR model fit to subjective happiness reports during this gambling mood induction paradigm computed mood based on fluctuations in concrete and token rewards, computed as a fading integration across the last ~7 trials (Fig 1E-G). To better understand the neural correlates of mood, we looked to determine how these behavioral signatures were reflected in high-resolution measures of AIC neural activity recorded in the rhesus monkey brain. Following on the behavioral observation that subjective happiness reflects an integration of

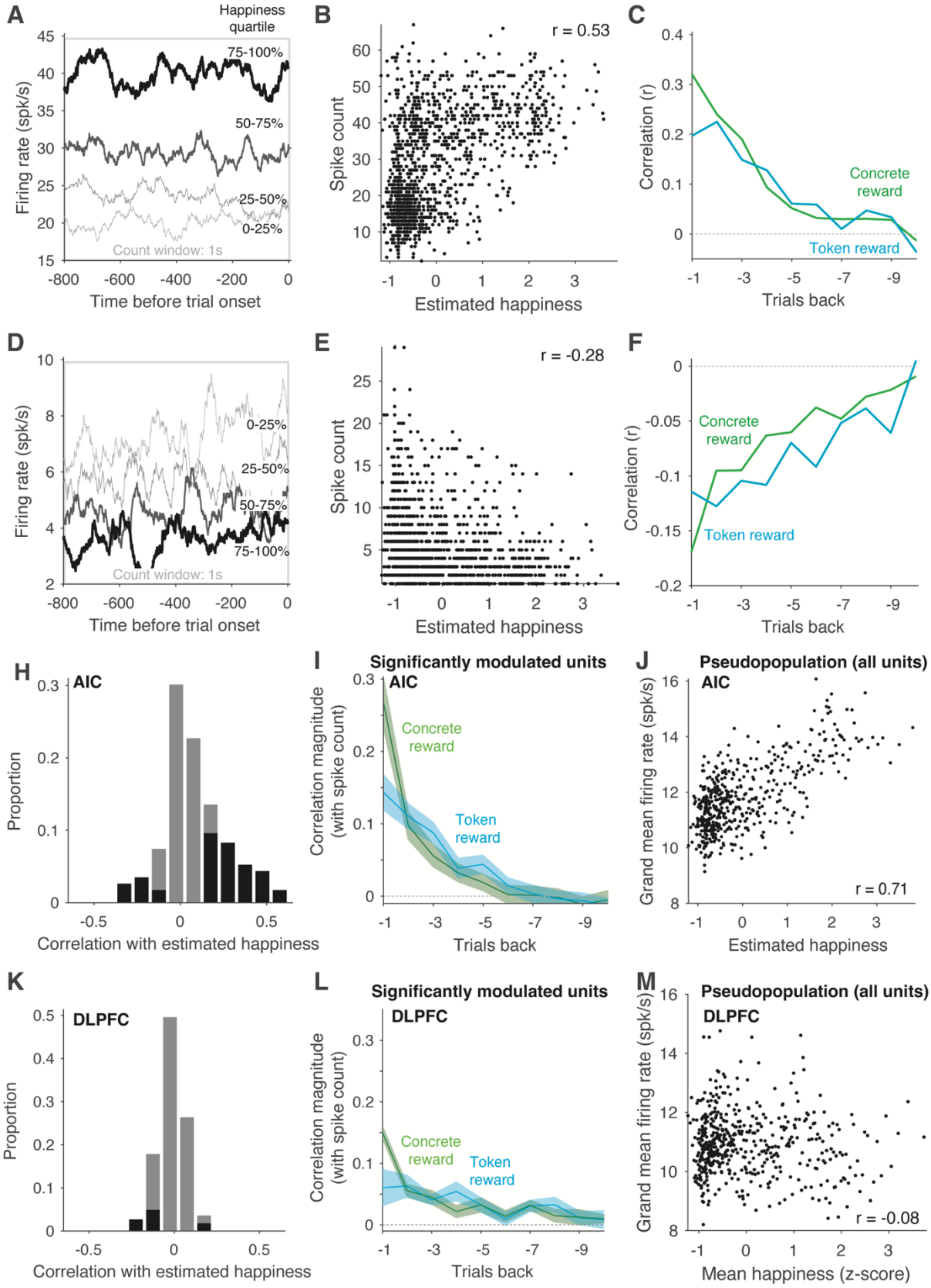

A
Happiness quartile
75-100%
50-75%
25-50%
0-25%
Count window: 1s
Firing rate (spk/s)
Time before trial onset
B
r = 0.53
Spike count
Estimated happiness
C
Concrete reward
Token reward
Correlation (r)
Trials back
D
0-25%
25-50%
50-75%
75-100%
Count window: 1s
Firing rate (spk/s)
Time before trial onset
E
r = -0.28
Spike count
Estimated happiness
F
Concrete reward
Token reward
Correlation (r)
Trials back
H
AIC
Proportion
Correlation with estimated happiness
I
Significantly modulated units
AIC
Concrete reward
Token reward
Correlation magnitude (with spike count)
Trials back
J
Pseudopopulation (all units)
AIC
r = 0.71
Grand mean firing rate (spk/s)
Estimated happiness
K
DLPFC
Proportion
Correlation with estimated happiness
L
Significantly modulated units
DLPFC
Concrete reward
Token reward
Correlation magnitude (with spike count)
Trials back
M
Pseudopopulation (all units)
DLPFC
r = -0.08
Grand mean firing rate (spk/s)
Mean happiness (z-score)

Figure 2. *Individual AIC units reflect estimated happiness.* A) A PSTH for one monkey AIC unit during the intertrial interval period in the 1s before trial onset, with trials grouped into estimated happiness quartiles. B) For the same unit, the correlation between estimated happiness based on the monkey's wins and losses and the spike count response where each dot reflects 1 intertrial interval period. This unit was positively correlated with estimated happiness. C) For the same unit, correlation between the spike count responses and both concrete and token reward history. D-F) The responses of a second AIC with responses negatively correlated with happiness, with the same conventions as panels A-C. H) Histogram of the Pearson correlation values (e.g., B, E) across all 235 neurons (gray) and those with significant correlations in black (significance criterion $p<0.001$). I) Correlation magnitudes (after taking the absolute value of the sign correlation) of spike counts with concrete and token rewards, averaged across all significant units. J) A pseudopopulation was created by aligning units by estimated happiness. Shown is the correlation between estimated happiness and AIC population response vigor. K-L) The same analyses described in panels H-J applied to 224 units recorded in DLPFC.

recent events, we suspected that mood might be reflected in AIC via the type of persistent, elevated activity reported in other brain regions where representations are often distributed across individual, persistent units (22).

To test this idea, we hypothesized that the monkey brain computes and reflects the same signatures that contribute to human subjective happiness reports, and one locus of that representation is AIC. Notably, our hypothesis is restricted to the nature of neural representations in the monkey brain (which we can measure) while remaining agnostic about whether the monkeys subjectively experience happiness (which we have no way to access). Because mood is an affective state that changes over a timescale much longer than a single trial, we reasoned that it should be reflected not just during gambling trials, but also between them. We thus probed the time period *in between* trials of the gambling task for mood representations.

To test this hypothesis, we computed estimated happiness by passing the monkeys' trial options and their outcomes through the universal mood model fit to humans. The responses of one remarkable AIC neuron are shown in Fig 2A, illustrating how the activity of this neuron persists throughout the 1s period before trial onset in a manner that reflects estimated happiness. Fig 3B shows a strong and highly significant correlation between estimated happiness and the firing rate of this individual unit between trials ($r=0.52$; $p=2*10^{-94}$), where every dot reflects the spike count in a single intertrial interval period. Fig 2C shows the correlation between this unit's firing and both types of reward (concrete and token) as a function of trial history, illustrating an integration over ~7 trials for concrete and token rewards (compare with 1D). In sum, during the period between trials of the gambling task, this unit reflected all the signatures of human subjective happiness in its persistent responses, including modulation by both concrete and token rewards, integrated across several trials. We found many units of this type (elaborated below).

While less frequent, we also found AIC units whose responses were strongly negatively correlated with estimated happiness. One example is shown in Fig 2D-F. Like the unit depicted in Fig 2A, this unit's firing rates persisted across the 1s period before trial onset (Fig 2D); however, the firing rate of his unit was lower when estimated happiness was higher ($r=-0.27$; $p=6*10^{-5}$; Fig 2D-E). Likewise, this unit's firing rates decreased based on concrete rewards and total token number across the last ~9 trials (Fig 2F).

Across the insula population, the intertrial interval responses of a notable 38% of insula units were significantly correlated with estimated happiness (significance criterion: $p<0.001$). When

parsed by positive versus negative correlations, 3.8-fold more units were positive (proportions 30% positive versus 8% negative of the total population, respectively; Fig 2H). Correlation magnitudes averaged across significantly modulated neurons (with the correlation signs flipped for negatively correlated happiness units) reflected an integration across ~6 trials back (Fig 2I), and bore a striking resemblance with behavior (compare with Fig 1D).

We also assessed how fluctuations of the combined AIC population aligned with the behavioral universal mood model. In these experiments, the neurons were recorded individually, most often in different sessions. We thus created a pseudopopulation by aligning trials by their estimated happiness similar to the approaches described by (23). The resulting pseudopopulation contained the responses of 229 AIC units on 480 trials. Correlations between estimated happiness and the overall vigor of the AIC population response were strong and highly significant r=0.71 ($p=1*10^{-73}$; Fig 2J) , and this was true in each monkey individually (monkey 1: r=0.66, $p=3*10^{-65}$; monkey 2: r=0.39, $p=2*10^{-18}$). The strength of these correlations is notable given that it rivals the strength of correlations reflected in visual cortex for visually-evoked responses (23) but in the case presented here for AIC, this happens during the behaviorally unconstrained intertrial interval period and the variable in question is based on subjective happiness reports.

We wondered whether mood might be reflected not just in AIC, but ubiquitously across the many higher brain areas demonstrated to reflect persistent activity. For insight, we applied the same analyses to units recorded in dorsolateral prefrontal cortex (DLPFC) and there we found a very different result. Only 9.4% of DLPFC units were significantly correlated with estimated happiness (significance criterion: $p<0.001$; Fig 2K; compared with 38% in AIC, Fig 2H). Among significant units, correlation magnitudes were weaker in DLPFC than in AIC (Fig 2L; compare with Fig 2I). When DLPFC units were configured into a pseudopopulation (the responses of 224 units on 480 trials), estimated happiness was very weakly negatively correlated with the overall vigor of the DLPFC population response (Fig 2M; r=-0.08; p=0.07). These results suggest that the behavioral signatures of mood are reflected much more strongly in AIC than in DLPFC and are thus not universal across all brain areas with persistent activity.

As mentioned above, the CRTR model was one of a class of several behavioral models we considered (Supplementary Table 2). The CRTR model estimated happiness based on four parameters that captured two types of events (model 4.1, concrete and token rewards, mean r^2 = 0.39). In comparison, one of the best models estimated happiness based on 6 parameters that captured concrete rewards and token-related events parsed into 3 terms (model 6.2, mean r^2 = 0.43). Those included 1 term that captured tokens received when the certain reward option was selected. In addition, 2 terms captured tokens received on gambling trials, including 1 to capture variations in the probability-weighted average (expected value) and a second to capture the difference between the expected value and the amount won (reward prediction error). Conversely, the worst model estimated happiness based on the total token number alone (model 3.3, mean r^2=0.09).

Variation in the average goodness of prediction that exists between these models (mean r^2=0.09 - 0.43) provides an opportunity to assess whether the factors that make for a better versus worse model of human subjective happiness are related to the factors that make for a better versus worse model of what drives the vigor of the persistent firing rates in AIC. We found that these

had a strong and significant relationship in which rank-order was largely preserved (Figure 3A, r=0.88; p=0.009). Intriguingly, this relationship appears nonlinear (Fig 3A, dashed gray line). These results are consistent with the hypothesis that human subjective happiness reports are reflected in monkey AIC as persistent activity.

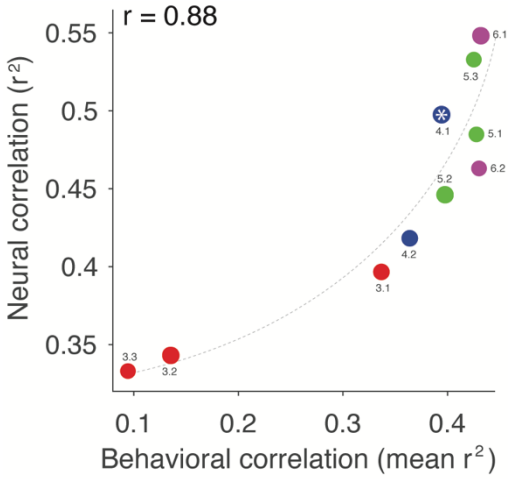


Figure 3. *Model comparison for behavioral versus neural data.* Plotted are the mean r^2 for each behavioral model (Supplementary Table 2) versus the correlation of each model with AIC neural data (e.g. Fig 2J) with each model labeled. The CRTR model (4.1) used in Fig 2 is labeled with an asterisk. Colors reflect the number of parameters in each model (3: red; 4: blue; 5: green; 6: magenta).The gray dashed line exists as a visual aid to illustrate the nonlinear relationship between these two quantities.

As a final set of analyses, we used the suite of mood models to investigate whether persistent activity in AIC reflected memories of past rewards as compared to the anticipation of future ones. Because the gambling paradigm involves the accumulation of tokens to receive concrete rewards, the two are correlated in this experiment (e.g. following a trial in which 3 tokens are won, persistent activity could reflect the token reward on the last trial or the anticipation that concrete reward may be received on the next one). However, following trials in which a concrete reward is delivered, 6 tokens are subtracted, and consequently, the probability of receiving a reward on the next trial is zero (because a concrete reward requires 6 tokens and the maximum that can be won is 3). As such, the ability of the concrete reward model alone (model 3.1) to account for the behavioral ($r^2 = 0.34$) and neural ($r^2 = 0.40$) data cannot be explained by reward anticipation; it can only be explained by reward memory. In comparison, a model based on reward anticipation alone loosely maps onto total token number (model 3.3), and of all the models we considered, this model provided the worst account of our behavioral (r^2=0.09) and neural (r^2=0.33) data. Thus, while we cannot rule out the possibility that reward anticipation contributes in some fashion to persistent activity in AIC, we can conclude it is not the primary driving force behind it.

Discussion

Because mood is a subjective experience that we have no external way to measure, studying how the brain supports it requires different approaches than studying many other brain functions (like

perception, memory, and decision making). Here, we took a nontraditional approach to the challenge of studying mood by creating universal models to predict how the happiness of humans will fluctuate as they engage in a gambling task (based on the wins and losses they experience), coupled with investigation of the neural correlates of these fluctuations in the monkey brain.

Here, we demonstrate a remarkably strong alignment between estimated happiness and the responses of units in AIC. There, a notable 38% of neurons were significantly modulated by estimated happiness, and estimated happiness was strongly correlated with an estimate of population response vigor (r=0.71). The strength of this alignment is notable given that it rivals that of visual cortex for image properties that also modulate the population response (23). Of note is that in AIC, this alignment happens during the behaviorally unconstrained intertrial interval period (as opposed to a visually-evoked response) and the behavior is shaped around subjective answers to the question, "How happy are you right now?". In addition, we found that both happiness reports (Fig 1D) and the firing rate of AIC units (Fig 1I) correlated with reward-related events across similar numbers of trials (~7). Furthermore, the factors that made for better versus worse behavioral models were correlated with those that accounted for AIC response vigor (Fig 3). In comparison, estimated happiness was only weakly reflected in DLPFC (Fig 2K-M), suggesting that our results do not follow from a signal (like arousal) broadcast across all higher brain areas. Together, these results suggest that persistent activity in AIC may be a neural correlate of the subjective experience of human happiness in the context of this gambling task. Even if this is not true (because the neural correlate lies elsewhere or involves AIC only in part), these results shed insight into how happiness is computed in the primate brain.

Namely, our results inform how the brain transforms positive and negative experiences into mood during the gambling mood induction paradigm, constrained by the following signatures. First, mood is reflected in AIC as persistent activity. Second, mood depends on both concrete and token rewards (with relative weightings of ~1/0.5). Third, the computation of mood integrates both types of events across the past ~7 trials, as a fading average. Fourth, in AIC the neural correlates of mood are dominated by neurons that positively correlate with happiness, but negatively correlated neurons exist as well (with ~4x more positively tuned units).

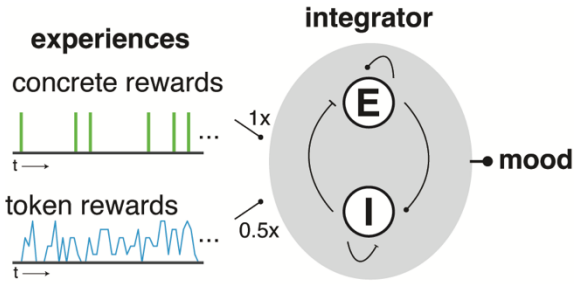


Figure 4. *An emerging model of how the brain transforms experiences into moods.* The fact that happiness reflects an integration of experiences combined with the finding that it is reflected in AIC as persistent activity is consistent with models in which integration happens via a population of reciprocally coupled excitatory and inhibitory neurons (24).

Together, these results are reminiscent of network models that perform integration and create persistent activity by operating via recurrently connected push-pull pools of positive and

negative neurons (Fig 4) (24). Here, we extend those ideas to sketch a model of how mood is computed in the brain — which may happen within AIC or may happen elsewhere and simply be reflected there. In this framework, mood is computed based on experiences of concrete and token rewards (with 1/0.5 relative weightings), which are fed into a recurrent network configured to integrate information across ~7 trials via reciprocally connected pools of excitatory (E) and inhibitory (I) units (Figure 4). The resulting network output then reflects mood as persistent activity, including population response vigor in AIC. This account (Fig 4) focuses on the computation of happiness as integrated reward (the CRTR model), for its demystifying conceptual simplicity. More complex but also more accurate accounts of reward-related events parse them into quantities such as the amount of reward that is expected (expected value) and surprising deviations from it (reward prediction error) (19, 25). A variant of the model presented in Figure 4 is one in which concrete and token rewards are replaced with these quantities.

Like the bulk of neurophysiology studies, the results we present here are correlative. Based upon them, we do not know whether persistent activity in AIC drives happiness or even if it is necessary for it. However, these results lay the foundation required to design causal perturbation experiments to investigate that question. Several other lines of evidence implicate AIC in the conscious experience of feelings including mood (20), including natural human lesions (26), fMRI (19, 27, 28), and microstimulation (29, 30). Our study complements this other work by providing insight into the nature of mood representations in AIC at the resolution of individual units and their action potentials. Intriguingly, attempts to decode mood from the human brain via intracranial electrocorticogram (ECoG) measures have implicated a host of other limbic brain areas outside of AIC in mood (such as cingulate and orbitofrontal cortex as well as the amygdala and hippocampus, (16)). Likewise, attempts to decode depression from human ECoG recordings have implicated overlapping brain areas including DLPFC (17), whereas we found that estimated happiness was only weakly reflected there in our study. Future work will be required to tease apart these differences to determine whether they follow from different mood induction paradigms (e.g. everyday mood fluctuations versus the gambling induction paradigm we use here), differences in spatial resolution of the recording techniques (single unit vs. ECoG), or species differences.

We found that a model of subjective happiness that integrates reward across trials provided a reasonable estimate of happiness fluctuations, and that this integration was reflected via persistent activity and population response vigor in AIC. Other studies have also found evidence for persistent representations of integrated reward in other brain areas, including anterior cingulate cortex (ACC, (31)) where (as in our study of AIC), the fraction of units of this type was significantly higher than DLPFC (32). These results are consistent with strong functional homologies between AIC and ACC, although AIC has been more strongly associated with emotional awareness (33). In comparison, persistent representations of reward in ACC have been linked to the ongoing evaluation of choice outcomes based on reward history in a game that requires continuous monitoring of strategy and adjustments of it (31). These two interpretations of the function for reward-related persistent activity are not mutually exclusive, as happiness not only reflects integrated experience but in turn, also impacts decision making (1, 34). Analogous persistent representations of reward across trials have also been demonstrated in rodents (35–37).

Likewise, while we have no way of determining what the monkeys experience as they perform this task, our approach is grounded in the subjective experience of humans, which may be essential for arriving at an impactful understanding of mood. For emotions such as fear, concerns have been raised that the circuits that mediate innate behaviors (including threat-related behaviors such as freezing) may be distinct from those that mediate subjective experience and consequently, the behaviors that mediate innate behavior may be of limited therapeutic relevance (14). By focusing on the neural correlates of human subjective experience, our approach seeks to bridge between neuroscience work in humans and animals using a model-based approach. Going forward, a parallel pursuit of all three approaches — work in humans, work in animals, and approaches that bridge across species— will be crucial for an impactful understanding of mood.

Methods:

As an overview, data from two sources are included in this paper. First, behavioral and neural data collected from monkeys performing a gambling task — including a study also reported here (21). That report examined the relationship between neural responses in anterior insular cortex (AIC) and decisions in a gambling task to test ideas related to the prospect theory of decision making. The results presented here focus exclusively on an epoch that was not analyzed in that report (the period between trials) and these results cannot be inferred from that paper. We also include unpublished data collected from the same two monkeys in the original report, but from dorsolateral prefrontal cortex (DLPFC).

In addition, this study includes behavioral data collected from humans performing the same gambling task with additional subjective happiness probes, inspired by the approaches reported by (19), with modifications.

The Methods associated with each type of data are described below.

Behavioral and neural data collected from monkeys performing a gambling task (absent happiness probes):

Because the methods used to collect this data are described in detail by (21), we summarize them here. All animal care and experimental procedures were conducted in accordance with the US public Health Service policy on the humane care and use of laboratory animals and were approved by the Johns Hopkins University Institutional Animal Care and Use Committee (IACUC).

Experiments were performed on two adult male rhesus macaque monkeys (Monkey G: 7.2 kg, 7 years; Monkey O: 9.5 kg, 8 years). All behavioral training and testing were performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. MonkeyLogic software (38) (https://www.brown.edu/Research/monkeylogic/) was used to control task events, stimuli, and reward, as well as monitor and store behavioral events.

We used T1 and T2 magnetic resonance images (MRIs) (3.0 T; Johns Hopkins Hospital) to determine the recording locations in both the anterior insular cortex (AIC) and dorsolateral prefrontal cortex (DLPFC) of each monkey. AIC, a distinct cortical lobe situated within the Sylvian (lateral) fissure, is implicated in interoception, emotion, and mood regulation. It is bordered anteriorly by the orbital prefrontal cortex and covered dorsally and ventrally by the frontoparietal and temporal opercula, respectively. DLPFC is located in the frontal lobe and involved in cognitive control and decision-making, was targeted near the principal sulcus, specifically in areas 8Ad and 8Av, slightly dorsal to the posterior principal sulcus. Recording sites in AIC and DLPFC were estimated using the stereotaxic coordinates of the recording chamber and electrode penetration depths, and superimposed onto MRI scans (Supplementary Figures 4 and 5).

Single neuron activities were recorded extracellularly with single tungsten microelectrodes using the TDT system (Tucker & Davis). Action potentials were amplified, filtered, and discriminated conventionally with a time–amplitude window discriminator. Neural data were recorded from anterior insular cortex, guided by stereoscopic chamber placement based on MRI images, and recording electrode depth.

Behavioral data collected from humans performing a gambling task (with happiness probes):

International, English-speaking subjects (n=55) were recruited through the online website Prolific and run on the online platform Pavlovia driven by PsychoPy code. Subjects were adults (age range 18-63, median 33) that self-identified as 47% male, 48% female, 5% undisclosed. The behavioral experiment was conducted with the approval of the University of Pennsylvania Institutional Review Board for the Protection of Human Subjects (IRB).

The token-based gambling task:

The token-based gambling task consisted of two types of trials: choice trails and forced choice trials. In choice trials, two targets (a sure option and a gamble option) were presented on the screen. In forced choice trials, a single target was presented. In both cases, the token values of the target(s) were indicated. For monkeys, the token values corresponding to options were indicated by colored boxes; for humans, numbers were displayed (Figure 1). Monkeys selected an option by making a saccade to it; humans selected an option with a button press.

Each trial began with a display of the token count, followed by the options and a selection. When the total token count was greater than 6, a reward was delivered at the end of the trial (a liquid reward for monkeys; a monetary reward of $0.50 for humans) and 6 tokens were subtracted from the total. Humans began the experiment with a $6.00 base reward.

Here, we focus primarily on the neural representation in the insula during the period between trials, which always exceeded 1.7 seconds. During this period, the monkeys' behavior, including their eye movements and location of focus, was unconstrained.

The total number of choice conditions was 24, including all possible combinations of 4 sure options and 3 gamble options (for a total of 12) in symmetric gain (win token) and loss (lose token) conditions (for a total of 24). In addition, each of the 13 options were presented alone on forced choice trials. The 37 trial types were presented in a pseudorandom block design in which each trial type was presented before re-randomizing the sequence. In the human experiments, 4 blocks in total were presented. In the monkeys, the number varied from session-to-session; the minimum in this dataset was 12.

For human subjects, a happiness probe was interleaved every 2-3 trials. On these trials, subjects were presented with the question, "How happy are you at this moment (1-20)?" Subjects entered their rating (from 1-20) with a button press. Subjects completed a practice session before data were collected. For quality control, subjects whose happiness reports were negligibly impacted by events during the gambling task were eliminated by applying a generous p-value criterion to the correlation between happiness reports and concrete rewards on the previous trial ($p<0.15$). With this screen, 11 (of 55) subjects were eliminated and 44 were retained for further analysis.

<u>Universal mood model:</u>

We compared a number of models to determine the one that provided a best account of the behavioral data (Supplementary Info). Following on Rutledge et al., every model had two parameters in common: a baseline happiness term w_0 and a parameter that determines the number of trials to integrate over (γ) in the equation below. In addition, each model had one or more parameters that corresponded to the weights w_n for each variable included in the model. The generic form of the model can be described as:

$$happiness(t) = w_0 + w_1 \sum_{j=1}^{t} \gamma^{t-j} X1_j + \ldots$$

where w_1 is the weight for the variable X1. For instance, the CRTR model used in the main body of the paper had 4 parameters (w_0-w_2, and γ):

$$happiness(t) = w_0 + w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j + w_2 \sum_{j=1}^{t} \gamma^{t-j} TR_j$$

Here, TR refers to the number of tokens won on each trial (token reward) and CR to whether a concrete reward was received on that trial (which happened when tokens reached 6).

The parameters were fit to the combined data across 43 human subject using the optimization toolbox in MATLAB. They were then used it to estimate happiness for the each of the 44th subject's held-out data. To predict happiness fluctuations during the experiment (as opposed to

the factors, possibly external, that contributed to the baseline), happiness measures for each subject were z-scored and the w_0 term was not used to make model predictions. To estimate happiness for the monkeys, we fit the universal mood model to all 44 human subjects' data.

Population-based analysis of neural data:

AIC neural data include 235 units, parsed into 137 neurons recorded in monkey G and 98 neurons recorded in monkey O. DLPFC data included 224 units, parsed into 144 units in monkey G and 100 in monkey O. Both data sets were collected as the animals performed the gambling task.

Accurate estimate of population response magnitude requires many units to eliminate the selectivity (sometimes called "tuning") individual sampled units. To perform our analyses, we concatenated units across sessions to create a pseudopopulation using methods analogous to those described by (23). When creating this pseudopopulation, we aligned data across sessions in a manner that preserved estimated happiness. The number of trials that could be included in the pseudopopulation was limited by the session for which the fewest trials were obtained across both monkeys. For the other sessions, a matched number of trials was subselected by ranking happiness for that session, preserving the lowest-ranked and highest-ranked happiness within that session, and selecting the number of additional images required as those with estimated happiness that were evenly spaced between the two extreme happiness estimates for that session.

The resulting pseudopopulation consisted of the responses on 480 trials. To perform the neural analyses (insert Fig 2J, Fig 2M, Fig 3), an estimated happiness for each of the 480 trials was computed as the mean of the estimated happiness scores that were aligned to produce the pseudopopulation response for that trial.

Acknowledgements: This work was supported by the Simons Foundation (Simons Collaboration on the Global Brain award 543033 to N.C.R. and a Pivot Fellowship to N.C.R.).

References:

1. D. Bennett, G. Davidson, Y. Niv, A model of mood as integrated advantage. *Psychol. Rev.* **129**, 513–541 (2022).

2. E. Bliss-Moreau, P. H. Rudebeck, Animal models of human mood. *Neurosci. Biobehav. Rev.* **120**, 574–582 (2021).

3. J. A. Russell, Core affect and the psychological construction of emotion. *Psychol. Rev.* **110**, 145–172 (2003).

4. W. N. Morris, *Mood* (Springer, 1989).

5. E. Eldar, R. B. Rutledge, R. J. Dolan, Y. Niv, Mood as Representation of Momentum. *Trends Cogn. Sci.* **20**, 15–24 (2016).

6. National Institutes of Mental Health, "Any Mood Disorder [fact sheet]- National Institute of Mental Health (NIMH)."

7. R. S. McIntyre, *et al.*, Treatment-resistant depression: definition, prevalence, detection, management, and investigational interventions. *World Psychiatry* **22**, 394–412 (2023).

8. K. S. Kendler, J. Campbell, Expanding the domain of the understandable in psychiatric illness: an updating of the Jasperian framework of explanation and understanding. *Psychol. Med.* **44**, 1–7 (2014).

9. G. E. Simon, N. Moise, D. C. Mohr, Management of Depression in Adults: A Review. *JAMA* **332**, 141–152 (2024).

10. S. Jauhar, *et al.*, A leaky umbrella has little value: evidence clearly indicates the serotonin system is implicated in depression. *Mol. Psychiatry* **28**, 3149–3152 (2023).

11. J. J. Taylor, H. G. Kurt, A. Anand, Resting State Functional Connectivity Biomarkers of Treatment Response in Mood Disorders: A Review. *Front. Psychiatry* **12**, 565136 (2021).

12. S. Reardon, Pressure grows to ditch controversial forced swim test in rodent studies of depression. (2024).

13. L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, S. D. Pollak, Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychol. Sci. Public Interest J. Am. Psychol. Soc.* **20**, 1–68 (2019).

14. V. Taschereau-Dumouchel, M. Michel, H. Lau, S. G. Hofmann, J. E. LeDoux, Putting the "mental" back in "mental disorders": a perspective from research on fear and anxiety. *Mol. Psychiatry* **27**, 1322 (2022).

15. O. Herreras, Local Field Potentials: Myths and Misunderstandings. *Front. Neural Circuits* **10**, 101 (2016).

16. O. G. Sani, *et al.*, Mood variations decoded from multi-site intracranial human brain activity. *Nat. Biotechnol.* **36**, 954–961 (2018).

17. J. Xiao, *et al.*, Decoding Depression Severity From Intracranial Neural Activity. *Biol. Psychiatry* **94**, 445–453 (2023).

18. S. Alagapan, *et al.*, Cingulate dynamics track depression recovery with deep brain stimulation. *Nature* **622**, 130–138 (2023).

19. R. B. Rutledge, N. Skandali, P. Dayan, R. J. Dolan, A computational and neural model of momentary subjective well-being. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12252–12257 (2014).

20. A. D. B. Craig, How do you feel--now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* **10**, 59–70 (2009).

21. Y.-P. Yang, X. Li, V. Stuphorn, Primate anterior insular cortex represents economic decision variables proposed by prospect theory. *Nat. Commun.* **13**, 717 (2022).

22. J. Zylberberg, B. W. Strowbridge, Mechanisms of Persistent Activity in Cortical Circuits: Possible Neural Substrates for Working Memory. *Annu. Rev. Neurosci.* **40**, 603–627 (2017).

23. A. Jaegle, *et al.*, Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife* **8**, e47596 (2019).

24. S. Lim, M. S. Goldman, Balanced cortical microcircuitry for maintaining information in working memory. *Nat. Neurosci.* **16**, 1306–1314 (2013).

25. H. Keren, *et al.*, The temporal representation of experience in subjective mood. *eLife* **10**, e62051 (2021).

26. A. Ibañez, E. Gleichgerrcht, F. Manes, Clinical effects of insular damage in humans. *Brain Struct. Funct.* **214**, 397–410 (2010).

27. F. Vinckier, L. Rigoux, D. Oudiette, M. Pessiglione, Neuro-computational account of how mood fluctuations arise and affect decision making. *Nat. Commun.* **9**, 1708 (2018).

28. I. Molnar-Szakacs, L. Q. Uddin, Anterior insula as a gatekeeper of executive control. *Neurosci. Biobehav. Rev.* **139**, 104736 (2022).

29. J. Yih, D. E. Beam, K. C. R. Fox, J. Parvizi, Intensity of affective experience is modulated by magnitude of intracranial electrical stimulation in human orbitofrontal, cingulate and insular cortices. *Soc. Cogn. Affect. Neurosci.* **14**, 339–351 (2019).

30. F. Bartolomei, *et al.*, The role of the dorsal anterior insula in ecstatic sensation revealed by direct electrical brain stimulation. *Brain Stimulat.* **12**, 1121–1126 (2019).

31. H. Seo, D. Lee, Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *J. Neurosci. Off. J. Soc. Neurosci.* **27**, 8366–8377 (2007).

32. H. Seo, D. Lee, Cortical mechanisms for reinforcement learning in competitive games. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **363**, 3845–3857 (2008).

33. X. Gu, P. R. Hof, K. J. Friston, J. Fan, Anterior Insular Cortex and Emotional Awareness. *J. Comp. Neurol.* **521**, 3371–3388 (2013).

34. E. Eldar, Y. Niv, Interaction between emotional state and learning underlies mood instability. *Nat. Commun.* **6**, 6149 (2015).

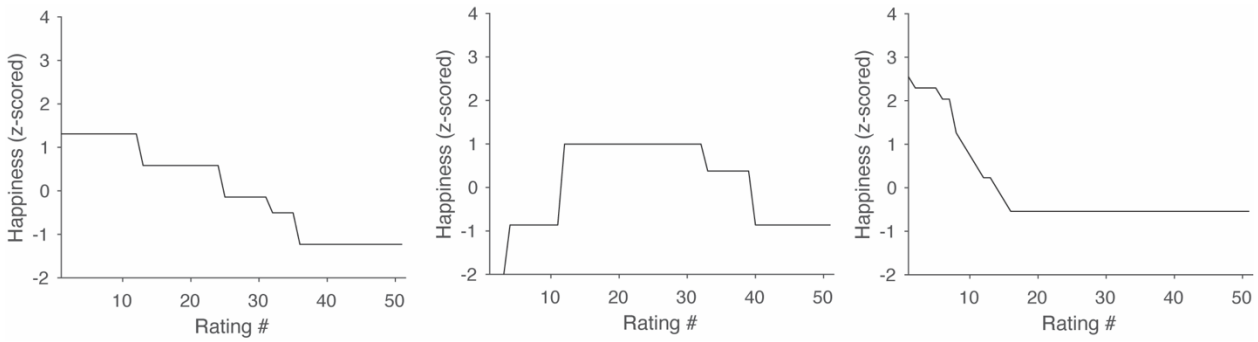
35. N. A. Steinmetz, P. Zatka-Haas, M. Carandini, K. D. Harris, Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273 (2019).

36. M. Song, *et al.*, Hierarchical gradients of multiple timescales in the mammalian forebrain. *Proc. Natl. Acad. Sci.* **121**, e2415695121 (2024).

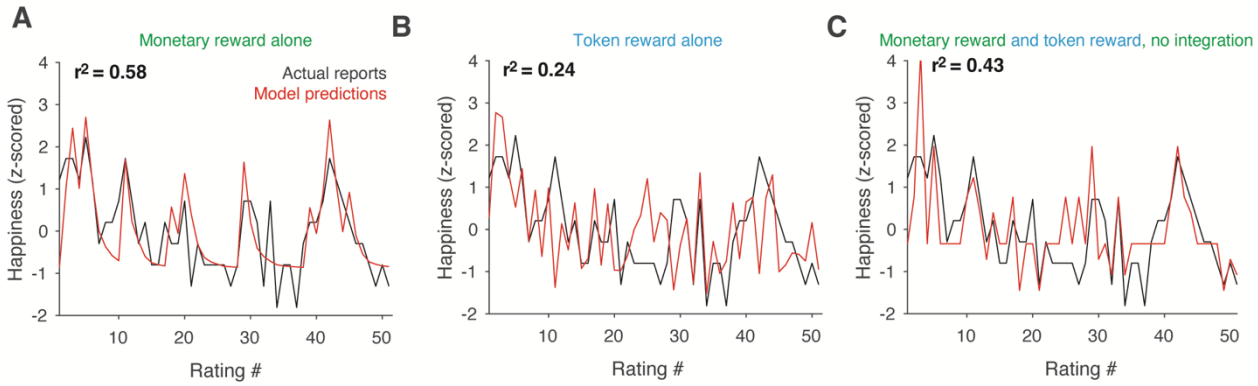
37. R. Hattori, T. Komiyama, Context-dependent persistency as a coding mechanism for robust and widely distributed value coding. *Neuron* **110**, 502-515.e11 (2022).

38. W. F. Asaad, E. N. Eskandar, A flexible software tool for temporally-precise behavioral control in Matlab. *J. Neurosci. Methods* **174**, 245–258 (2008).

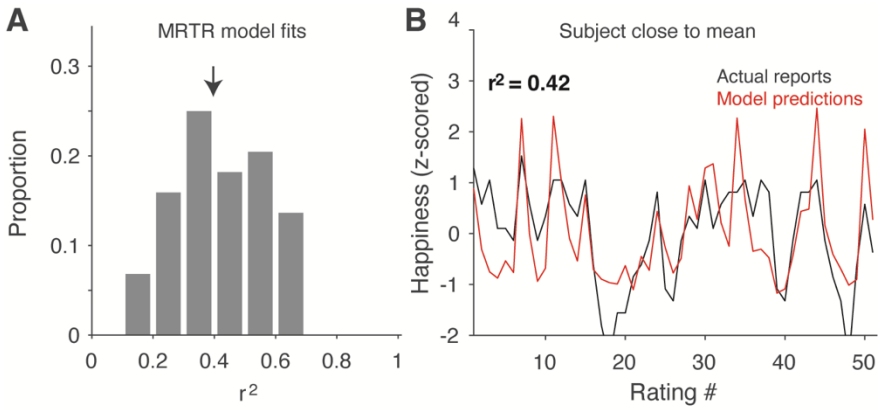
Supplemental information



Supplemental Figure 1. *Illustration of 3 subjects whose subjective happiness reports were insensitive to the events that happened during the gambling mood induction paradigm.* Compare these traces to the traces for the subject depicted in Supplemental Figure 2 (black). In these 3 subjects, happiness fluctuates along a slow timescale that either decreases across the experiment (for the leftmost and rightmost subject, possibly due to boredom) or gradually rises and then falls (middle subject). For 11 (of 55) subjects (including these 3), subjective happiness reports were not significantly correlated with concrete rewards on the previous trial with a generous p-value criterion ($p<0.15$) and these subjects were removed from further analyses.



Supplemental Figure 2. *Model predictions for the subject shown in Fig 1E ($r^2 = 0.71$) with simplified versions the CRTR model.* In all panels, the same actual happiness reports for this subject (black) are compared with different model predictions (red) where model predictions were created by fitting the model to 43 subjects to predict the 44th subject's held out data. **A)** Model based on integrated concrete reward alone (Model 3.1 in Table 2). **B)** Model based on integrated token reward alone (Model 3.2 in Table 2). **C)** Model based on combined concrete and token reward with out integration (and thus just the events on the previous trial). For this subject, all simplified versions of the CRTR model provided worse predictions of the data.



Supplemental Figure 3. *CRTR model fits.* **A)** Distribution of r^2 across all 44 subjects, where where model predictions were created by fitting the model to 43 subjects to predict this 44th subject's held out data. Arrow indicates mean (0.39). **B)** Illustration of predictions for the subject closest to the mean, showing that the model captures main features of this individual's subjective happiness fluctuations.

Comparison of behavioral mood models:

In the study that inspired the behavioral component of this work, Rutledge et al. (1) compared the ability of a class of models to predict individual subjects' happiness fluctuations as they won and lost money on each trial of a gambling task. In comparison, the gambling task used to collect the neural component of this work (Yang et al. (2)) differed in a notable way: subjects played for abstract, token rewards and received concrete ones (money for humans; juice for monkeys) when the number of tokens reached six. Here we thus must revisit the question of which model within the general class Rutledge et al. (1) considered accounts for happiness fluctuations with these modifications.

Our goal is also a bit different than the original Rutledge study insofar as we were focused on creating a model that captured universal fluctuations in happiness (as opposed to individual differences). To achieve this, we fit data combined across 43 subjects to predict happiness fluctuations for each of the 44th subject's data. To predict happiness fluctuations during the experiment (as opposed to the factors, possibly external, that contributed to the baseline), happiness measures for each subject were z-scored and the w_0 term (described below) was not used to make model predictions. We measure goodness of prediction as the Pearson r^2 between actual reports and held out model predictions.

Below we present these models by introducing their simplest versions first and then building them up to incorporate additional parameters. Following on Rutledge et al., every model has two parameters in common: a baseline happiness term w_0 and a parameter that determines the number of trials to integrate over (γ) in the equation below. In addition, each model had one or more parameters that corresponded to the weights w_n for each variable included in the model. For instance, the 3 parameter model that computed happiness based only on concrete reward (CR) took the form:

$$happiness(t) = w_0 + w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j$$

Similarly, the 4 parameter model that computed happiness based on concrete rewards (CR) and token rewards (TR) took the form:

$$happiness(t) = w_0 + w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j + w_2 \sum_{j=1}^{t} \gamma^{t-j} TR_j$$

A description of all variables we considered can be found in Table 1. The parameterization of all the models that we considered and their predictive power is summarized in Table 2.

CR	Concrete reward	Took on a value of 1 when a concrete reward was received and 0 otherwise.
TR	Token reward	Number of tokens won per trial; range -3 to 3
TT	Total token number	Number of total tokens at trial end; range 0 to 6
TS	Token start	Number of total tokens at trial start; range 0 to 5
CrR	Certain rewards	Number of tokens won on trials when the certain reward option was selected (nan otherwise); range -3 to 3
GR	Gamble rewards	Number of tokens won on trials when gamble option was selected (nan otherwise); range -3 to 3
EV1	Expected value (definition 1)	When gamble option was selected, probability weighted average between options (nan otherwise), as described by (1).
RPE1	Reward prediction error (definition 1)	When gamble option was selected, difference between tokens won and EV1, as described by (1).
EV2	Expected value (definition 2)	Average value of all token rewards received up to the trial in question (for both certain reward and gamble selections), as described by (3)
RPE2	Reward prediction error (definition 2)	Difference between tokens won and EV2, (for both certain reward and gamble selections), as described by (3)
EV3	Expected value (definition 3)	Average value of all token rewards received up to the trial in question, computed in the same manner as EV2 but only for gamble selections, as described by (3)
RPE3	Reward prediction error (definition 3)	Difference between tokens won and EV3, only for gamble selections, as described by (3)

Table 1. A description of all variables considered in the model fits.

Number of parameters	Model name	Events	mean r^2 all subjects
3	3.1	CR	0.34
3	3.2	TR	0.14
3	3.3	TT	0.09
4	4.1 (CRTR)	CR + TR	0.39
4	4.2	CR + TT	0.36
5	5.1	CR + CrR + GR	0.43
5	5.2	CR+ TR + TS	0.40
5	5.3	CR + EV2 + RPE2	0.42
6	6.1	CR + EV3 + RPE3+ CrR	0.43
6	6.2	CR + EV1 + RPE1+ CrR	0.43

Table 2. Mean r^2 between subjective happiness reports and model predictions, computed for 44 subjects. Model 4.1 (CRTR) was used to create figure 2.

Models based on single events (3 parameters): The best fitting model of the single-event class computed happiness based on concrete rewards (CR; model 3.1, mean r^2=0.34). Equivalent models based on token rewards (TR, model 3.2, mean r^2=0.14) or total token number (TT, model 3.3, mean r^2=0.09) provided worse accounts of the data.

These results suggest that (unsurprisingly), concrete rewards are the singular biggest contributor to subjective happiness. Across all the models we considered (despite their number of parameters), models that did not include a concrete reward term provided worse predictions. To avoid overwhelm and instead focus on evaluating the models that provide the best fits, we do not discuss models without concrete reward further.

Models based on concrete rewards and token events (4 parameters): This class of models focused on accounting for concrete rewards and 1 parameter that captured some aspect of token-related events. The best fitting model of this class incorporated both concrete rewards (CR) and token rewards (TR, model 4.1, mean r^2=0.39). In the main text, we call this the "CRTR model". An equivalent model based on CR and total token number provided a slightly worse account of the data (TT, model 4.2, mean r^2=0.36).

When compared with 3 parameter models, these results suggest that both concrete and token rewards contribute to happiness fluctuations.

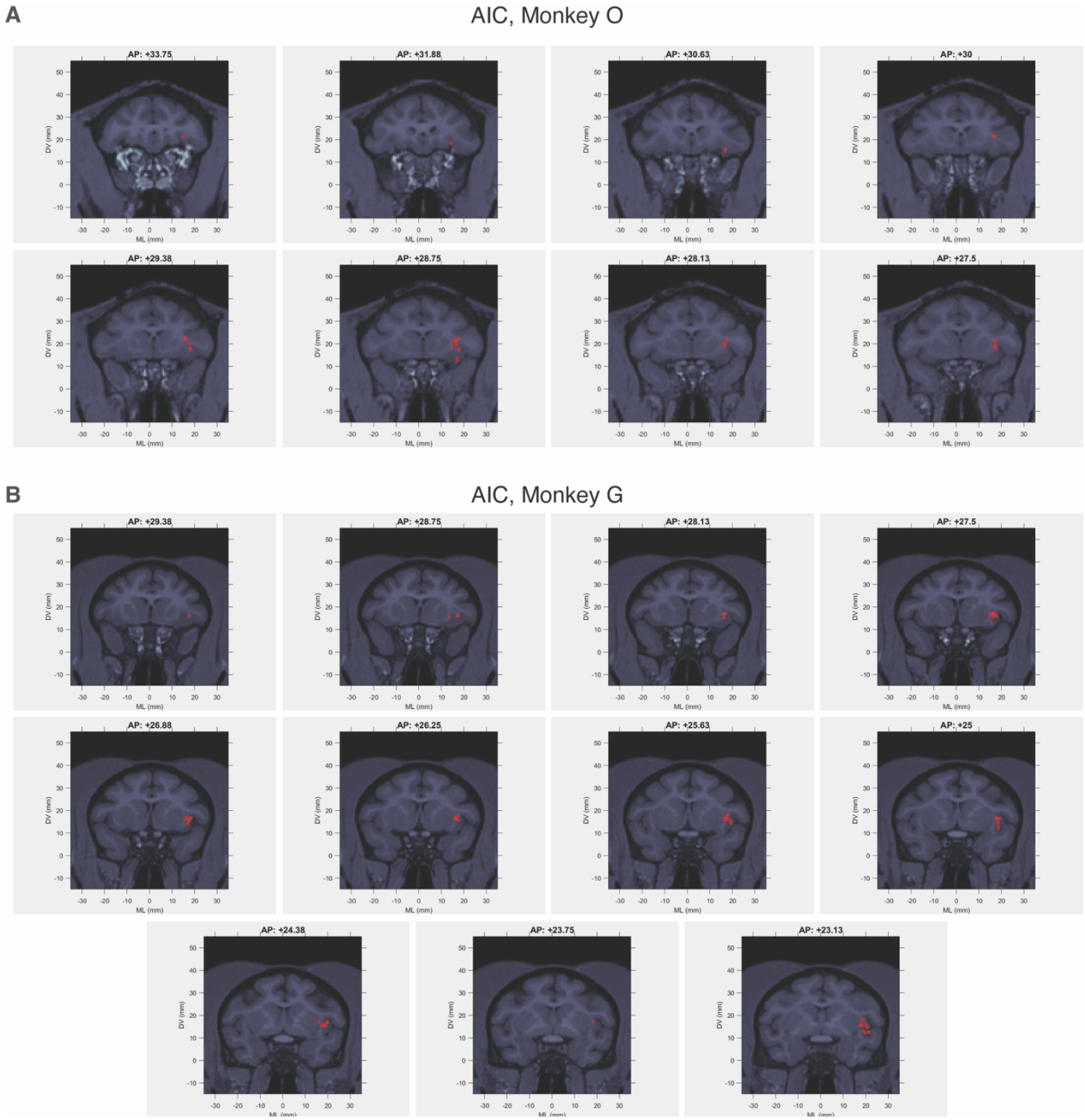
Models based on concrete rewards and a decomposition of token-related events into 2 terms (5 parameters): This class of models broke down token-related events into two terms, formulated in different ways. One best fitting model of this class parsed token events into tokens received when the certain reward (CrR) versus the gambling (GR) option was selected (as suggested by (1), model 5.1, mean r^2=0.43). Similar fits were obtained when tokens were decomposed into expected value (EV2) and reward prediction error (RPE2), computed as described by (3), where EV2 was computed as the average of all token rewards up to the trial in question and RPE2 was computed as the difference between the tokens received and that value

(model 5.3, mean mean r^2=0.42). In comparison, a model that decomposed the total number of tokens into token reward (TR) and tokens at the start of each trial (TS) did not perform quite as well (model 5.2, mean r^2=0.40).

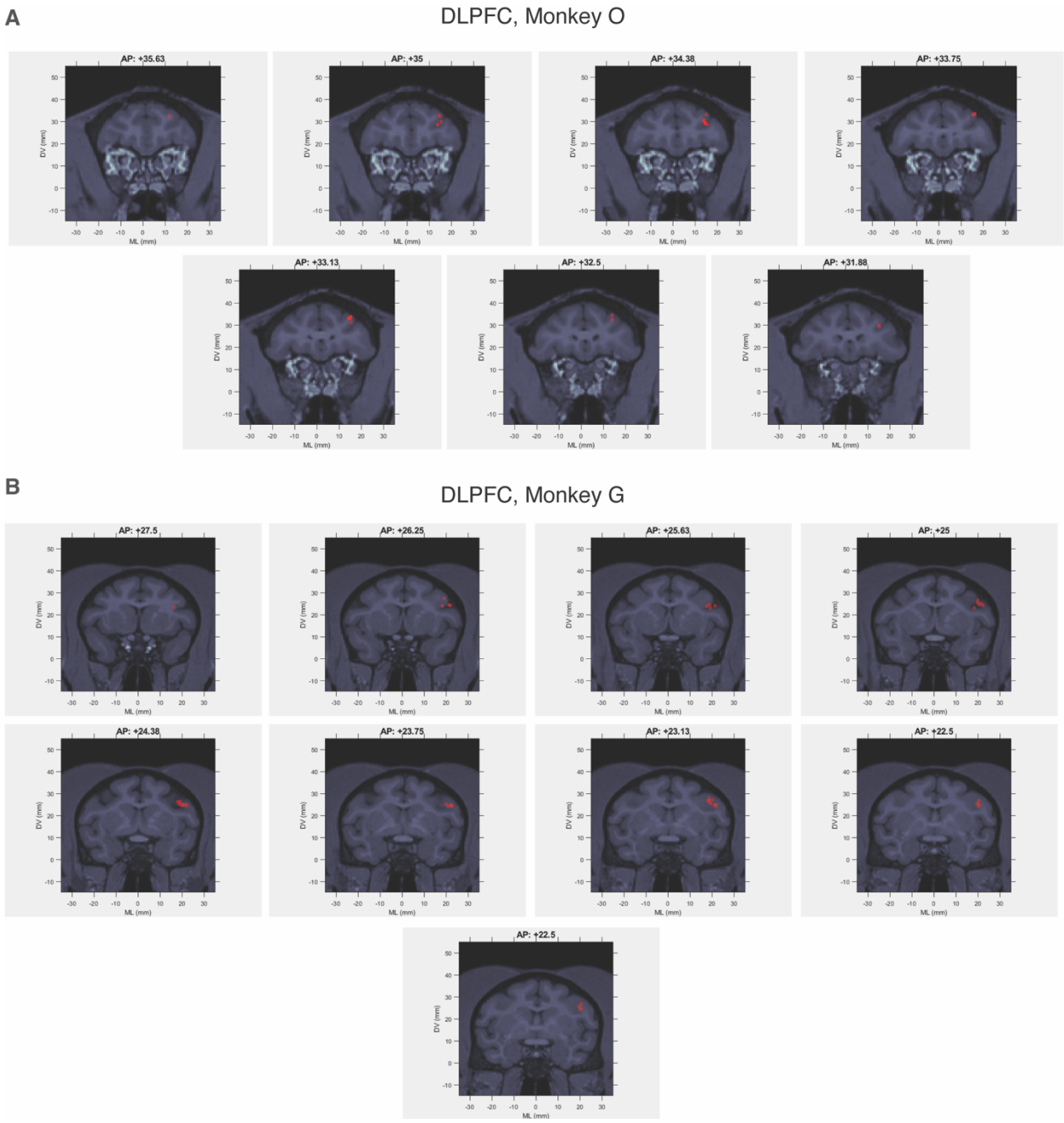
These results suggest that happiness depends not just generically on the numbers of tokens won or lost on each trial. Rather, models that allow for happiness to be differentially weighted when (e.g.) the tokens won on each trial were won when the certain reward versus the gambling option was selected provide better fits.

Models based on concrete rewards and a decomposition of token-related events into 3 terms (6 parameters): This class of models focused on accounting for concrete rewards and token events parsed into 3 terms, including 1 term corresponding to tokens received when the certain reward (CrR) option was selected and 2 terms corresponding to tokens received on gambling trials. Events on gambling trials parsed two ways produced equivalent model fits. The first defined expected value (EV3) similar to model 5.3: as the average outcome up to each trial in question but limited in this case to gambling trials, following on ((3), model 6.1, mean r^2=0.43). The second defined expected value (EV1) as the average between the two options on gambling trials, following on ((1), model 6.2, mean r^2=0.43). In both cases, reward prediction error (RPE3 and RPE1) was computed as the difference between the gambling trial outcome and EV.

General conclusions. While the 6 parameter models provided the best fits overall (models 6.1 and 6.2, mean r^2 = 0.43), the best 5 parameter models perform nearly as well (model 5.1 mean r^2 = 0.42). Likewise, performance of the best fitting 4 parameter model is slightly but not markedly less (model 4.1, mean r^2 = 0.39). Given the similarity of model fits, coupled with the similarities in the parameterization of these models, we hesitate to declare that there is a singular "best model" that accounts for subjective happiness reports in this data. Rather, a more parsimonious description is that a class of models in which happiness depends on a weighted combination of concrete rewards and token-related events is a decent description of the data overall, and breaking token-related events into event subtypes generally accounts for additional variance (with some nuances). For the clarity of presentation in the main body of the paper, we describe our approach to the neural data by focusing on model 4.1 (where happiness depends on concrete and token reward, mean r^2 = 0.39) due to the conceptual simplicity of that model. We then follow with an exploration of how those results generalize to other, more complex models (Figure 3).



Supplemental Figure 4. *AIC recording locations.* **A)** Monkey O. **B)** Monkey G.



Supplemental Figure 5. *DLPFC recording locations.* **A)** Monkey O. **B)** Monkey G.

References:

1. R. B. Rutledge, N. Skandali, P. Dayan, R. J. Dolan, A computational and neural model of momentary subjective well-being. *Proc Natl Acad Sci U S A* **111**, 12252–12257 (2014).

2. Y.-P. Yang, X. Li, V. Stuphorn, Primate anterior insular cortex represents economic decision variables proposed by prospect theory. *Nat Commun* **13**, 717 (2022).

3. H. Keren, *et al.*, The temporal representation of experience in subjective mood. *eLife* **10**, e62051 (2021).