Subjective emotion judgements adhere to principles of Bayesian inference and efficient representation

Jade R Serfaty[1] and Quentin JM Huys[1]

[1*]Applied Computational Psychiatry Lab, Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology and Mental Health Neuroscience Department, Division of Psychiatry, University College London, London, UK.

Contributing authors: jade.serfaty.17@ucl.ac.uk; q.huys@ucl.ac.uk;

Abstract

Emotion judgements are central to wellbeing and mental health, yet the fundamental principles determining how they arise remain poorly characterised. A key challenge is that repeated emotion ratings show substantial variability, typically treated as noise. Here we propose that emotion judgements, like perceptual judgements, arise from probabilistic inference under representational constraints. Using behavioural and computational approaches, we show that this variability is structured and meaningful. Emotion judgements are biased toward prior expectations when evidence is weak, consistent with Bayesian inference, and representational precision scales with the frequency of experienced intensities, consistent with efficient coding. Confidence decreases as uncertainty increases, indicating partial awareness of this uncertainty. A single model combining efficient encoding with Bayesian decoding captures these behavioural signatures. Notably, these signatures were largely preserved across variation in anxiety symptoms. These findings link emotion self-report to the principles of perceptual inference, providing a computational account of how uncertainty structures emotion judgements.

Keywords: emotion judgements, efficient coding, Bayesian inference, uncertainty, confidence, anxiety

Summary

Emotion judgements are central to everyday social life, basic research and mental health, yet the mechanisms by which people judge and report their own emotional states remain poorly understood. A key challenge is that repeated emotion ratings show robust variability, typically dismissed as noise, although this variability may instead reflect uncertainty in an underlying inferential process. Constructed and appraisal theories of emotion imply a role for such uncertainty, but this has not been formally tested in emotion judgements [1, 2]. In sensory perception, uncertainty is captured by Bayesian inference and efficient coding [3, 4]. Here we show that these same principles apply to emotion judgements. Across five independent samples, including a pre-registered replication, repeated emotion ratings predicted downstream behaviour, reported intensities were drawn toward prior expectations when evidence was weak, and discriminability was enhanced for more frequently experienced intensities. Confidence further tracked this uncertainty, and a single computational model combining efficient encoding with Bayesian decoding captured the resulting variability, bias and discriminability. These findings suggest that emotion judgements obey the same core principles that govern perceptual inference, and provide a formal framework for analysing emotional inference in health and disease.

1 Introduction

Emotion judgements shape many aspects of society, influencing how we interact with others and make decisions, and their objective measurement plays a core role in the diagnosis and treatment of mental illness. Yet, the fundamental cognitive processes that govern emotion judgements remain poorly understood [5–15]. A central challenge lies in the conflict between the assumption that emotion judgements provide direct access to a 'true' internal emotion intensity and the empirical reality that repeated ratings in emotion measurements show substantial variability [16]. When we ask someone how they feel, we typically assume their answer reflects that internal truth, and treat their report as authoritative - if I say I feel happy right now, what grounds do you have to disagree? Yet such variability reveals that emotion judgements are far less stable than this assumption implies.

This tension is reflected in prominent emotion theories, which offer contrasting views on the nature of emotion and emotion judgements. Basic-emotion accounts treat emotions as discrete and innate, with stable signatures across contexts, leaving little room for graded variation [17, 18]. Other accounts emphasise the integration of multiple sources of information, though they differ in how this process is formalised. Appraisal theories view emotions as evaluations of goal relevance, implying context-dependent integration [2]. Constructed accounts emphasise context-sensitive categorisation from core affect and conceptual knowledge [1, 9]. Componential models foreground coordination across appraisal, physiological, action and feeling components [19], and neural perspectives emphasise distributed population codes integrating interoceptive and exteroceptive signals [12].

Despite their differences, these perspectives converge on an inferential view: emotion judgements arise from integrating multiple, noisy sources of information, such that variability may reflect uncertainty in this inferential process rather than mere measurement noise.

We build on this idea by proposing that emotion judgements constitute a form of perceptual inference. As in perception, the true intensity of an emotion elicited by a stimulus is not directly accessible; the brain must infer it from noisy internal signals, yielding estimates that vary with context and task demands. This inference depends on how prior knowledge is weighted against incoming sensory evidence and how representational precision is allocated. In sensory domains, these computations are well captured by Bayesian and efficient coding principles, providing a potential framework for understanding variability in emotion judgements. Efficient coding formalises how limited neural resources are allocated in proportion to the statistics of encountered values in the environment [3, 20–22]. These principles have also been extended beyond sensory perception to subjective value-based decision-making [4].

If emotion judgements are indeed perceptual in this sense, they should exhibit the same computational signatures observed in other perceptual domains. This framework raises several questions about how emotion judgements operate, which we tested across multiple emotions. First, is the variability observed in emotion judgements meaningful rather than noise, and does it predict downstream behaviour? Second, do emotion judgements follow Bayesian inference, such that estimates are drawn toward prior expectations when evidence is weak? Third, is representational precision allocated according to efficient coding principles, with higher precision near frequently experienced intensities? Finally, does confidence reflect reliability, decreasing when repeated ratings are more variable?

We then asked whether a single computational model combining efficient encoding with Bayesian decoding could jointly account for these patterns. These hypotheses were tested in three independent samples and evaluated for robustness in a larger, pre-registered replication. In the replication, we additionally conducted exploratory analyses examining how the mechanisms proposed in our model of emotion judgements relate to symptoms of anxiety, given the central role of emotion judgements in psychopathology and its diagnosis. Finally, we ran a fifth study in a group of participants reporting anxiety symptoms to test these hypotheses with greater power.

2 Results

2.1 Variability in emotion judgements shapes choices

Emotion judgements are known to vary across observations. However, what remains unclear is whether this variability is mere measurement noise added after the underlying 'true' emotion has been identified and judged, or whether it reflects a more meaningful feature of the internal processes or representations that generate the emotion judgement itself. If the variability were only noise, its associated uncertainty would be irrelevant when combining or comparing judgements. If, instead, the variability arises within the generative process, the uncertainty should propagate to any behaviour that relies on these judgements, as is routinely observed

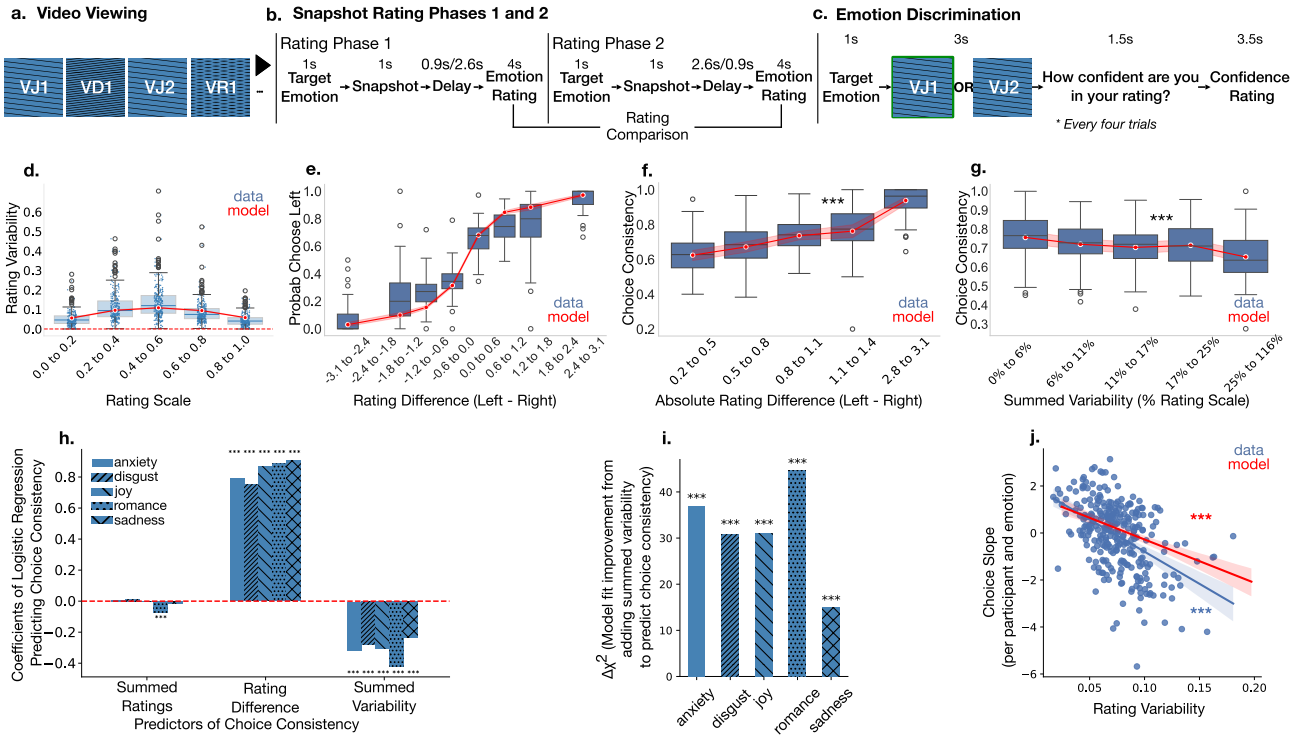

Fig. 1 Variability in Emotion Judgements Predicts Choice Behaviour. **a. Video Viewing.** Participants (Study 1, N = 57) viewed 142 brief video clips selected to elicit joy, romance, anxiety, disgust or sadness, in a random order, while performing an unrelated attention task. **b. Snapshot Rating.** On each trial, participants were shown a target emotion (e.g. "Joy") and then a snapshot clearly identifying one of the videos viewed previously. They then had 0.9s or 2.6s introspection time followed by 4s to enter the rating of the target emotion on a continuous scale. The snapshot rating phase was repeated a second time without prior warning, resulting in two ratings for each snapshot, one with 0.9s, the other with 2.6s introspection time. **c. Emotion Discrimination.** After rating all snapshots twice, participants chose which of two clips would be most helpful to induce a target emotional state in themselves. They were presented with two snapshots of the videos they previously rated, and were asked to choose the one that would best set them in the mood for the target emotion. Every fourth choice, they also reported their confidence. **d. Rating variability.** Snapshot ratings between the two phases per participant and emotion varied at all levels of intensity. Medium average ratings elicited more variable ratings than snapshots with high or low average ratings. **e. Choices.** The left snapshot was chosen more frequently when it had higher relative average rating on the target emotion. **f. Choice consistency and rating difference.** Choice consistency increased with increasing difference between the mean ratings of the snapshots. **g. Choice consistency and rating variability.** Choice consistency decreased with increasing variability of snapshot ratings. **h. Predicting choice at trial level.** Logistic regression coefficients predicting choice consistency from summed ratings (the sum of both options' mean emotion intensity ratings), summed variability (the sum of both options' standard deviations of emotion intensity ratings), and rating difference (the absolute difference between the options' mean emotion intensity ratings). **i. Effects of adding rating variability to predict choice.** Adding summed variability to predict choice consistency improves model fit ($\Delta\chi^2$) for every emotion category. Bars reflect the improvement in model fit from nested model comparisons. **j. Predicting choice at participant level.** Mixed-effects logistic regression slopes (effect of rating difference on choice consistency) plotted against rating variability per participant-emotion pair. The higher the participant's overall rating variability for an emotion, the weaker their sensitivity to rating differences for that emotion. Throughout, boxplot centre lines show medians; boxes show 25th–75th percentiles; whiskers extend 1.5× the IQR; outliers shown as points. Asterisks indicate statistical significance (*** $p < 0.001$). Where shown, blue indicates empirical data and red indicates results from analyses repeated on data generated by the efficient coding model fit to Study 1 rating and choice data.

with perceptual phenomena, within domains such as vision, audition, and proprioception [23–25], and when integrating information across modalities such as vision and proprioception [26, 27]. To test which account holds, we first quantified the variability in participants' emotion judgements and then asked whether that variability carried functional consequences.

Participants in Study 1 (N=57) watched validated emotional video clips eliciting joy (n=30 videos), romance (n=27 videos), anxiety (n=30 videos), disgust (n=28 videos) and sadness (n=27 videos; Fig. 1a). They then rated one snapshot from each clip on a continuous scale, judging how strongly the video from which the snapshot was taken elicited a specific target emotion (Fig. 1b). After providing one rating for all snapshots, participants were asked, without prior warning, to provide a second rating for each snapshot. Participant-level rating distributions revealed substantial heterogeneity across individuals for all emotions (Supplementary Figs. S1-S5). As expected, emotion-intensity ratings differed across the two rating phases, with greater variability at medium than at extreme intensities (Fig. 1d).

To test whether this variability had functional importance, we next asked participants to use their judgements of emotional intensity to make a choice. If the variability observed in Fig. 1d arose purely from a process downstream of the evaluation itself, it should not have affected how those intensity judgements were used in subsequent choices, because the true judged intensity would have been internally known. Choices would then have been independent of rating variability. By contrast, if rating variability reflected a meaningful internal process or representation, it should also have influenced choices in a manner consistent with that uncertainty. We therefore asked participants to imagine themselves as actors having to put themselves into a specific emotional state. On each trial, they were presented with two previously rated snapshots and asked to choose which video would better help them achieve that emotional state if they viewed it again (Fig. 1c). The assumption was that participants should choose the snapshot judged as eliciting a stronger intensity of the target emotion, because it should be more effective. Choices followed a sigmoidal function of the mean rating difference between the two options (Fig. 1e). Moreover, choices were less consistent when the two options had more similar mean ratings (mixed-effects logistic regression: $OR = 2.54$, 95% CI $[2.30, 2.81]$, $p < 0.001$, Fig. 1f).

Consistent with perceptual decision-making accounts, choices were also faster when evidence was stronger, a pattern classically observed in perceptual judgements, where stronger stimulus evidence leads to faster and more accurate decisions [28–30]. Across all five emotion categories, consistent choices were associated with shorter log reaction times than inconsistent choices (linear mixed-effects models with random intercepts for participant: anxiety, $\beta = -0.020$, 95% CI $[-0.039, -0.002]$, $p = 0.032$; disgust, $\beta = -0.044$, 95% CI $[-0.064, -0.024]$, $p < 0.001$; joy, $\beta = -0.039$, 95% CI $[-0.058, -0.020]$, $p < 0.001$; romance, $\beta = -0.034$, 95% CI $[-0.054, -0.014]$, $p = 0.001$; sadness, $\beta = -0.031$, 95% CI $[-0.050, -0.011]$, $p = 0.002$; Supplementary Fig. S6a), and log reaction times decreased as the difference in mean ratings between the two options increased (anxiety, $\beta = -0.029$, 95% CI $[-0.046, -0.012]$, $p = 0.001$; disgust, $\beta = -0.041$, 95% CI $[-0.058, -0.023]$, $p < 0.001$; joy, $\beta = -0.044$, 95% CI $[-0.062, -0.027]$, $p < 0.001$; romance, $\beta = -0.050$, 95% CI $[-0.069, -0.030]$, $p < 0.001$; sadness, $\beta = -0.036$, 95% CI $[-0.051, -0.020]$, $p < 0.001$; Supplementary Fig. S6b).

If uncertainty were an integral part of the process giving rise to emotion judgements, it should also be reflected in less deterministic preferences between snapshots. Indeed, choice consistency decreased as the summed variability of the two choice options increased (mixed-effects logistic regression: $OR = 0.77$, 95% CI $[0.73, 0.82]$, $p < 0.001$, Fig. 1g). Furthermore, this effect should be visible within each emotion category and should be additional to, and opposite from, the impact of mean rating differences. For each of the five emotion categories tested, choices were most strongly predicted by the difference in average ratings between the two video snapshots (separate logistic regressions per emotion: joy, $\beta = 0.874 \pm 0.078$; sadness, $\beta = 0.911 \pm 0.08$; anxiety, $\beta = 0.794 \pm 0.08$; disgust, $\beta = 0.758 \pm 0.07$; romance, $\beta = 0.893 \pm 0.09$; all $p < 0.001$; Fig. 1h), with larger rating differences leading to more consistent choices. However, including the summed variability of the two snapshots significantly improved prediction beyond rating difference alone in every emotion category (joy: $\Delta\chi^2 = 31.00$; sadness: $\Delta\chi^2 = 14.98$; anxiety: $\Delta\chi^2 = 37.02$; disgust: $\Delta\chi^2 = 30.92$; romance: $\Delta\chi^2 = 44.79$; all df $= 1$ and $p < 0.001$; Fig. 1i). Examining the regression coefficients showed that, in every emotion category, greater variability reduced choice consistency (joy: $\beta = -0.307 \pm 0.05$; sadness: $\beta = -0.233 \pm 0.06$; anxiety: $\beta = -0.318 \pm 0.05$; disgust: $\beta = -0.281 \pm 0.05$; romance: $\beta = -0.422 \pm 0.06$; all $p < 0.001$; Fig. 1h). By contrast, the sum of mean intensity ratings across the two snapshots had no significant effect (Fig. 1h).

Finally, if variability reflects uncertainty in the underlying judgement process, then individuals with higher overall variability for a given emotion should also show reduced sensitivity to rating differences along that emotion. Consistent with this prediction, average rating variability negatively predicted the random slope of rating difference on choice consistency per emotion in a mixed-effects logistic regression model ($\beta_{\text{robust}} = -27.47 \pm 2.98$, $p < 0.001$; Fig. 1j).

These results replicated in an independent sample (N=120), in which the design was identical but limited to anxiety (Supplementary Fig. S10a). Replication analyses confirmed all key findings: variability peaked at intermediate intensities, mean rating differences predicted choices ($\beta = 0.85 \pm 0.04$, $p < 0.001$), variability reduced choice consistency ($\beta = -0.33 \pm 0.04$, $p < 0.001$) beyond rating differences ($\Delta\chi^2 = 80.66$, df $= 1$, $p < 0.001$), and participants with higher overall variability showed weaker sensitivity to rating differences ($\beta = -32.73 \pm 8.21$, $p < 0.001$; Supplementary Fig. S10b-h).

Thus, variability in emotional intensity ratings, both within and across participants, had functional consequences for decisions about emotion. These findings suggest that uncertainty does not arise downstream of emotion judgements as mere measurement noise, but instead reflects a meaningful property of the internal processes and representations that generate them.

2.2 Emotion judgements behave in keeping with Bayesian Inference

Uncertainty in emotion judgements therefore appears to reflect a meaningful property of the internal processes that generate them, rather than mere measurement noise. A natural next question is whether this uncertainty is managed according to the same computational principles that govern perceptual inference. As in perception, participants do not directly know the "true" emotion intensity elicited by a video clip and must infer it from noisy internal representations. In sensory domains such as vision, audition and proprioception, uncertainty is classically formalised within a Bayesian framework, in which prior expectations are integrated with noisy evidence in proportion to their relative uncertainty. If emotion judgements rely on similar inferential processes, they should show the same hallmark signature: when sensory evidence is weak (i.e., noisier internal representations), judgements should be drawn toward prior expectations, whereas when evidence is strong, sensory evidence dominates, reducing prior-driven bias (Fig. 2a).

Experimentally, real-world priors are unknown *a priori*. However, within an experimental session, participants can acquire expectations that reflect the distribution of experienced stimuli [31, 32]. In our task, the initial viewing phase shaped participants' expectations about the range of emotion intensities, from which we estimated individual priors using ratings across phases 1 and 2.

Because the video sets in Studies 1 and 2 were uniformly distributed in intensity (30 clips per emotion), they produced near-flat priors. At the participant level, rating distributions were relatively broad and weakly structured (Supplementary Figs. S1–S5), and when aggregated across participants this yielded only weak prior-driven bias (Supplementary Figs. S8 and S9). In Study 3, we induced a non-uniform prior by running the task in emotion blocks (anxiety then joy; N=80 clips each) and over-sampling moderate intensities (Fig. 2b).

Following the framework introduced by Polanía and colleagues [4], shorter stimulus exposures were expected to yield noisier internal representations, because participants had less time to construct an estimate of the emotion elicited by the stimulus. This should increase reliance on prior expectations. Each snapshot was therefore rated twice: once after a short exposure and once after a long exposure (order counterbalanced across the rating phases). Long-exposure ratings ($\hat{e}_{\sigma_{low}}$) served as the lower-noise reference, whereas short-exposure ratings ($\hat{e}_{\sigma_{high}}$) reflected a higher-noise internal representation. We quantified prior-driven bias as the deviation of short-exposure ratings from long-exposure ratings for the same video:

$$(\hat{e}_{\sigma_{high}} - e_0) - (\hat{e}_{\sigma_{low}} - e_0) = \hat{e}_{\sigma_{high}} - \hat{e}_{\sigma_{low}}$$

If emotion judgements reflect structured Bayesian inference, short-exposure ratings should be pulled toward the prior peak, and this pull should be strongest for stimuli whose long-exposure ratings lie far from that peak.

Aggregated short- and long-exposure rating distributions (z-scored per participant with their mode so that the prior peak aligns at $x = 0$) illustrate the learned prior for each emotion (Fig. 2c,d, left panels). The short–long difference (Fig. 2c,d, right panels) reveals a clear attraction toward the prior peak for both anxiety and joy when ratings were far from the peak (linear regression of bias on long-exposure ratings; left of prior peak: anxiety, $\beta = -0.24$, 95% CI $[-0.28, -0.19]$, $p < 0.001$; joy, $\beta = -0.24$, 95% CI $[-0.30, -0.18]$, $p < 0.001$; right of peak: anxiety, $\beta = -0.08$, 95% CI $[-0.11, -0.04]$, $p < 0.001$; joy, $\beta = -0.16$, 95% CI $[-0.19, -0.14]$, $p < 0.001$). We observed the same pattern in the replication sample (left of prior peak: $\beta = -0.15$, 95% CI $[-0.17, -0.13]$, $p < 0.001$; right of peak: $\beta = -0.14$, 95% CI $[-0.16, -0.11]$, $p < 0.001$; Supplementary Fig. S11b). Greater uncertainty amplified prior-driven bias, drawing ratings closer to the prior's peak.

A further Bayesian prediction is that stronger priors should exert greater pull on uncertain judgements. Accordingly, participants with narrower priors showed steeper attraction slopes, for both anxiety (linear regression across participants: $\beta = 0.32$, $p < 0.001$; Fig. 2e) and joy ($\beta = 0.32$, $p = 0.004$; Fig. 2f). We observed the same pattern in the replication sample ($\beta = 0.46$, $p < 0.001$; Supplementary Fig. S11c).

As a control analysis, we asked whether prior attraction was modulated by the amount of information available in the snapshot itself. We reasoned that when a snapshot clearly identified the original video clip, emotion judgements should rely more on the available sensory evidence from the snapshot and less on prior expectations, while snapshots without clear information would rely more on memory. Consistent with this account, informative snapshots from video clips eliciting anxiety showed a weaker attraction effect than less informative snapshots, reflected in a significant interaction between long-exposure rating and snapshot informativeness in a mixed-effects model ($\beta = 0.039$, 95% CI $[0.006, 0.073]$, $p = 0.022$) and this effect was replicated in the replication study ($\beta = 0.024$, 95% CI $[0.004, 0.045]$, $p = 0.021$). Thus, Bayesian attraction was attenuated when the snapshot itself provided more evidence about the underlying video clip.

Together, these findings indicate that uncertainty in emotion judgements shapes behaviour in a manner consistent with Bayesian inference, revealing that emotion self-assessments incorporate prior expectations much like perceptual estimates do.

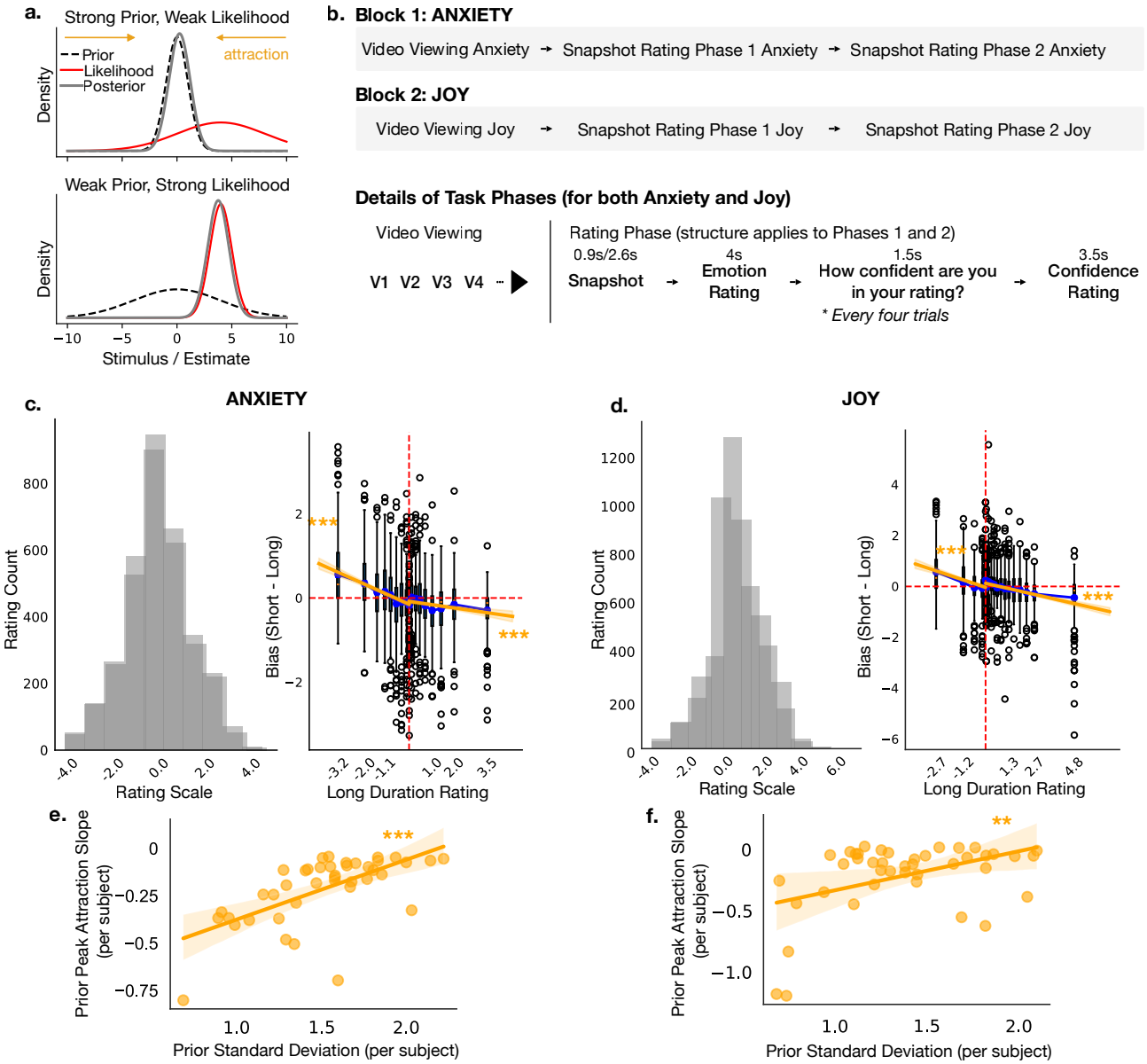

Fig. 2 Emotion Judgements Behave in Keeping with Bayesian Inference. **a. Bayesian inference illustration.** The posterior estimate (solid grey) results from combining a prior (black dashed line) and a likelihood (solid red). Top: when the prior is strong and the likelihood weak, the posterior is pulled toward the prior, demonstrating attraction. Bottom: when the likelihood is strong and the prior weak, the posterior remains closer to the observed evidence. **b. Task structure.** Study 3 (N=47) was run in two emotion blocks, anxiety and joy. In each block, participants first viewed videos and then completed two snapshot rating phases. Across the two rating phases, each snapshot was shown for a short (0.9s) and a long (2.6s) exposure time, followed by an emotion rating and, every four trials, a confidence rating. **c,d. Prior distributions and Bayesian attraction.** Left panels: aggregated rating distributions across the two rating phases for anxiety (**c**) and joy (**d**), z-scored per participant so that the participant-specific mode, used as the prior peak, is aligned at $x = 0$. Right panels: bias in emotion intensity judgements quantified as short–long rating difference (the "pull" toward the prior) plotted against the participant's z-scored long-exposure rating (the less-noisy measure), with each participant's mode aligned at x=0, for anxiety (N=41) and joy (N=42). The vertical dashed line marks the prior peak and the horizontal dashed line marks no difference between short- and long-duration ratings. Boxplots reflect the distribution of short–long differences binned by x-axis ratings (hinges = 25th/75th percentiles; whiskers = ±1.5 IQR; dots = outliers). Blue lines link the mean bias per bin (±s.e.m.). Overlaid piecewise linear fits (±95% CI) demonstrate that short-exposure ratings are pulled closer to the participant's prior peak when stimuli are far from that peak (orange). **e,f. Bayesian attraction and prior width.** OLS slopes from regressions of bias on long-duration ratings are plotted against participants' prior standard deviation for anxiety (**e**) and joy (**f**). Each point represents one participant. Participants with wider priors showed less negative attraction slopes, indicating weaker pull toward the prior peak. Asterisks indicate statistical significance (** $p < 0.01$, *** $p < 0.001$).

2.3 Emotion judgements behave in keeping with Efficient Coding

If emotion judgements reflect perceptual inference, they should not only integrate priors and evidence in a Bayesian manner but also allocate representational precision efficiently across possible intensities. Efficient coding theory proposes that neural systems devote more precision to values encountered most often. Applied to emotions, this predicts that intensities near the peak of a person's prior are represented with greater precision, shaping both bias in ratings and discrimination in choices.

These assumptions lead to clear behavioural predictions. Greater precision near the prior peak narrows and skews the likelihood function, creating longer tails away from the peak (Fig. 3a). This asymmetry yields a repulsive bias: ratings for stimuli just above or below the peak are pushed outward (Fig. 3b; see [3] for details). Discrimination performance should also be highest near the prior peak, where encoding precision is greatest (Fig. 3c,d), and repulsion should strengthen when internal noise increases (e.g., short exposures).

To quantify how encoding precision varied across the intensity range, we extended the Bayesian framework with an efficient coding model linking each video's underlying "true" emotion intensity (e_0) to an internal noisy neural response r (encoding) and its inferred estimate $\hat{e}(r)$ (decoding) (Fig. 3a). Encoding precision was modelled as a function of each participant's prior about the emotion intensity $p(e)$. With multiple ratings per stimulus, the model inferred the most likely underlying intensity under efficient coding assumptions. Simulations reproduced the predicted repulsive bias near the prior peak, amplified under higher internal noise (short exposure duration; Fig. 3e,f).

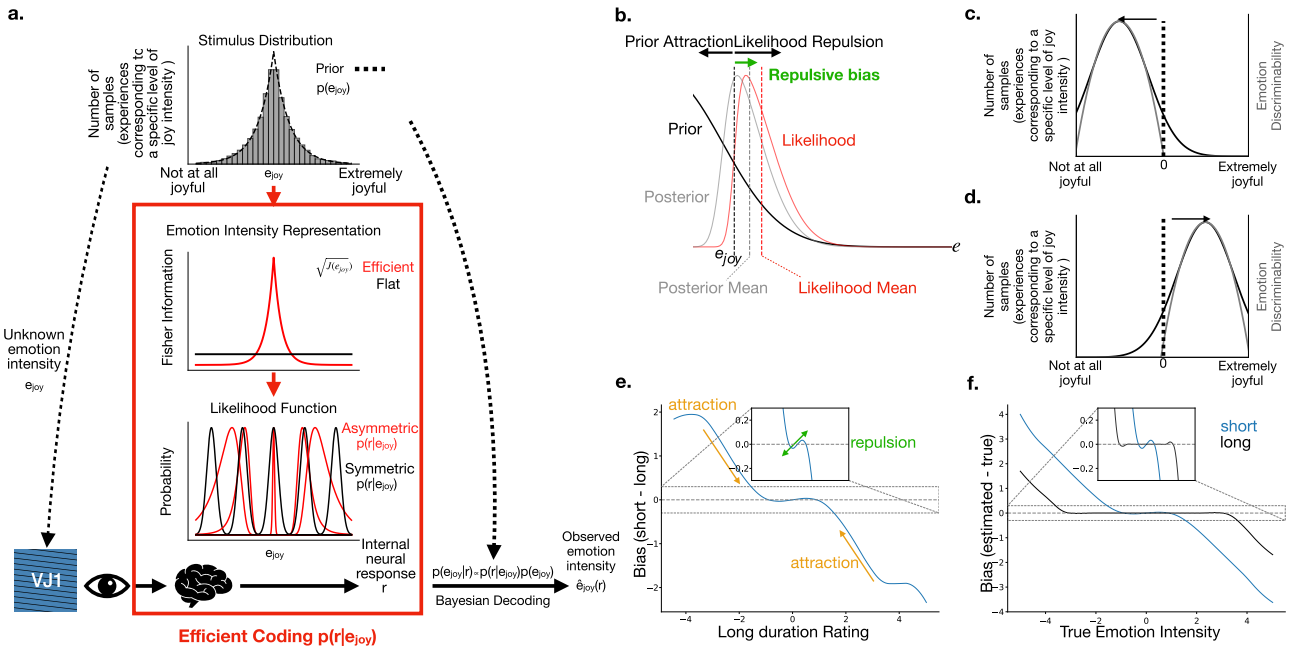

Fig. 3 Efficient Coding Framework and Behavioural Predictions. **a. Schematic of efficient encoding and Bayesian decoding of emotion judgements for joy.** A video with true (unobservable) emotion intensity e_{joy} elicits an internal, noisy neural response r, shaped by the observer's prior p(e_{joy}). Encoding is assumed to maximize mutual information between p(e_{joy}) and r, yielding higher encoding precision near the prior peak, and asymmetric likelihoods p(r|e_{joy}). Bayesian decoding combines likelihood and prior to generate an estimated emotion judgement ê(r), which is then reported behaviourally. **b. Efficient coding predicts a repulsive bias near the prior peak.** Near the prior peak, an asymmetric likelihood (red) can push estimates away from the peak (repulsion), partially countering the prior's attractive pull (black). **c,d. Emotion Discriminability and Prior.** Emotion discriminability (grey) is predicted to be greater where the prior (black) has higher density, whether that is at lower (**c**) or higher (**d**) intensities. **e. Simulated bias expressed as the difference between short and long exposure ratings.** Attraction at the prior tails reflects predicted Bayesian inference effects. The inset plot zooms into the prior peak region, highlighting predicted efficient coding effects: repulsion near the prior peak. **f. Simulated bias expressed as the difference between estimated and true emotion intensity.** The bias is shown for long (black) and short (blue) exposure durations. The inset plot provides a zoomed-in view around the prior peak, illustrating increased repulsion due to higher internal noise in the short exposure condition.

We tested these predictions empirically in Study 3. Bias was defined as the short-long rating difference in mode-aligned ratings. For both anxiety and joy, bias showed the predicted repulsive pattern near the prior peak (linear regression of bias near the peak: anxiety, $\beta = 0.55$, 95% CI $[0.12, 0.98]$ $p = 0.013$; joy, $\beta = 1.06$, 95% CI

[0.44, 1.68], $p < 0.001$; Fig. 4a,b). This repulsive pattern is a more specific signature of non-uniform encoding precision near the prior peak, rather than of prior bias alone.

If efficient coding confers greater representational precision near the prior peak, it should also improve discriminability between nearby video clips. This effect should be most apparent for difficult choices, that is, trials in which the two options have similar mean emotion ratings. We therefore restricted the analysis in the emotion discrimination task to small rating-difference trials and asked whether choices were more consistent when the choice-pair midpoint lay near rather than far from the prior peak. Consistent with efficient coding, stimuli closer to the prior peak yielded higher choice consistency than those farther away (Mann-Whitney $U = 532503.0$, $p < 0.001$; Fig. 4c), and within-subject analyses confirmed this effect (paired t-test: $t = 2.06$, $df = 51$, $p = 0.04$, Cohen's $d = 0.29$). The same relationship was observed when distance from the prior peak was treated continuously and choice difficulty, that is, rating difference, was controlled for: choice consistency decreased as the choice-pair midpoint moved farther from the prior peak (logistic mixed-effects model with random intercepts and random slopes for prior distance by participant: OR = 0.86, 95% CI [0.81, 0.92], $p < 0.001$; Fig. S12). As expected, choices were also more consistent when the rating difference between options was larger (OR = 2.30, 95% CI for OR [2.15, 2.47], $p < 0.001$).

We observed the same pattern in the replication sample: repulsive bias near the prior peak again replicated ($\beta = 0.67$, 95% CI [0.17, 1.17], $p = 0.009$; Supplementary Fig. S13a), and within-subject choice consistency was higher for stimuli closer to the prior peak (paired t-test: $t = 2.30$, $df = 27$, $p = 0.03$, Cohen's $d = 0.44$; Supplementary Fig. S13b).

Together, these results indicate that prior expectations shape not only bias, but also the precision of emotion judgement representations, consistent with efficient coding principles observed in sensory systems.

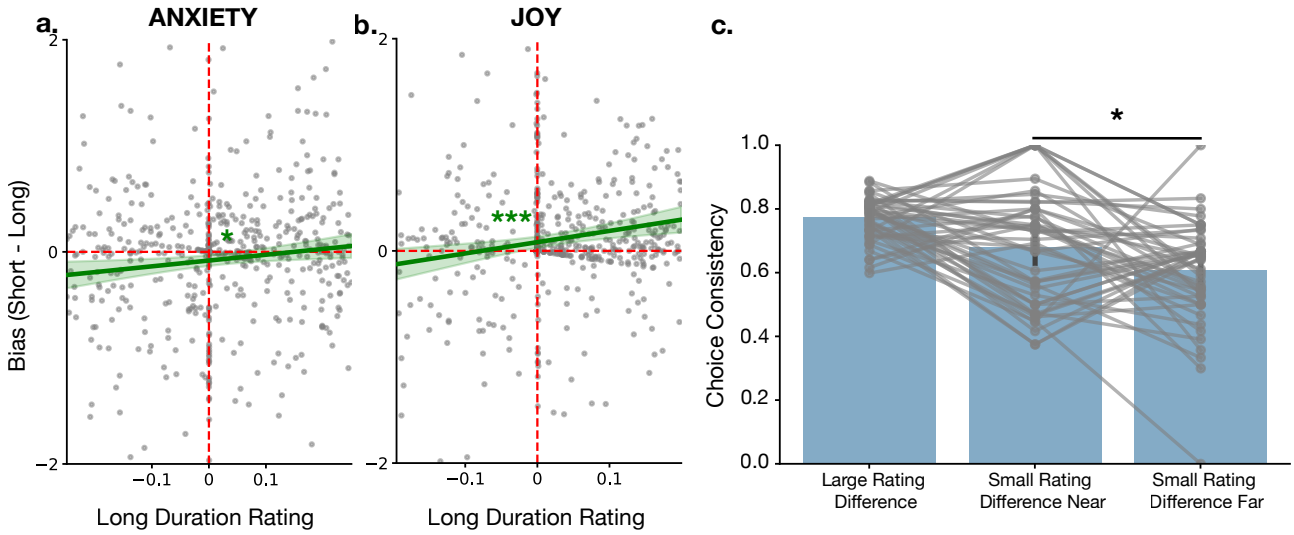

Fig. 4 Emotion Judgements Exhibit Systematic Biases Consistent with Efficient Coding. **a,b. Bias in emotion intensity judgements close to the prior peak.** Rating differences between short- and long-duration judgements are plotted against z-scored long-duration ratings, with each participant's mode aligned at zero, for anxiety (Study 3, N=41; **a**) and joy (Study 3, N=42; **b**). Green lines show linear fits ± 95% CI within the region close to the prior peak, highlighting a repulsive bias consistent with efficient coding. Red dashed lines indicate zero bias and the participant-specific prior peak. **c. Choice consistency near and far from the prior peak.** Choice consistency is shown for trials with small rating differences far from the prior peak, small rating differences near the prior peak, and large rating differences (Study 1, N=57). Grey lines link within-subject means; bars show group means ± 95% CI. Participants were more consistent for small-difference choices made near the prior peak than far from it (Mann-Whitney $p < 0.001$; paired t-test $p = 0.02$).

2.4 Awareness of uncertainty in emotion judgements

Our results thus far indicate that variability in emotion judgements arises from principled encoding and decoding processes, similarly to sensory perception. We next asked whether participants have insight into the reliability of their own emotion judgements, specifically whether confidence tracks uncertainty in our paradigm. If confidence tracks internal uncertainty, it should decrease when repeated ratings of the same snapshot are more variable. In choices, confidence should increase with evidence strength, that is, rating difference, but decrease with the summed variability of the two options.

To test whether participants tracked their own uncertainty, we collected confidence ratings for emotion intensity judgements in Study 3 (Fig. 2a). Rating variability was negatively associated with confidence after controlling for emotion ratings (mixed-effects model: β = -0.14, 95% CI $[-0.21, -0.08]$, $p < 0.001$; Fig. 5a). This pattern was replicated in the independent sample (β = -0.30, 95% CI $[-0.36, -0.25]$, $p < 0.001$, Supplementary Fig. S14a).

We next examined confidence in the emotion discrimination phase of Study 1 (Fig. 1c). Confidence increased with rating difference (β = 0.32, 95% CI $[0.26, 0.38]$, $p < 0.001$) and decreased more weakly with summed variability (β = -0.07, 95% CI $[-0.13, -0.01]$, $p = 0.03$; mixed-effects; Fig. 5b). Although the effect of rating difference on confidence was replicated (β = 0.32,95% CI $[0.27, 0.37]$, $p < 0.001$), the effect of summed variability was not (Supplementary Fig. S14b).

These findings show that confidence was sensitive to uncertainty in emotion judgements. Confidence in ratings decreased with rating variability, whereas confidence in choices was most strongly related to rating difference, with a weaker and non-replicating effect of summed variability.

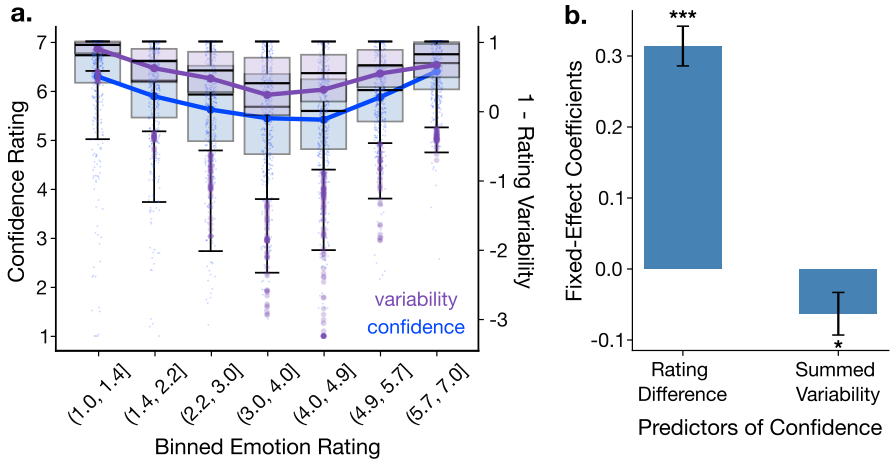

Fig. 5 Awareness of Uncertainty in Emotion Judgements. **a. Confidence in emotion intensity ratings and rating variability.** Confidence ratings (blue; left axis) and rating reliability, quantified as 1−rating variability (purple; right axis), are plotted against binned emotion intensity ratings per participant (Study 3, N=47). For confidence, boxplots show the distribution across bins, overlaid with individual observations, and solid lines indicate the mean in each bin. For rating reliability, boxplots show the distribution across bins, outliers are shown individually in light purple, and solid lines indicate the mean in each bin. In all boxplots, the lower and upper hinges correspond to the 25th and 75th percentiles, and whiskers extend to 1.5× IQR. **b. Predicting confidence in choices.** Fixed-effect coefficients from a linear mixed-effects regression predicting confidence in choices in the emotion discrimination task from rating difference and summed rating variability. A random intercept was included for each participant and emotion (Study 1, N=57). Error bars indicate the standard error of the estimate. ***$p < 0.001$; *$p < 0.05$.

2.5 Mathematical Model of Efficient Representation of Emotion Judgements

Having established that emotion judgements exhibit both Bayesian attraction and efficient-coding signatures, we next asked whether these effects could arise from a single underlying computational mechanism. Specifically, we tested whether a unified model representing uncertainty explicitly in both encoding and decoding could quantitatively reproduce the observed behavioural patterns.

The model builds on the framework introduced above (see Methods 3.9). Each video clip has a latent true emotion intensity (e_0), which is encoded through a monotonic transformation $F(e)$ derived from the participant-specific prior $p(e)$, then perturbed by internal noise (σ_{int}). The decoded estimate $\hat{e} = F^{-1}(r)$ is then perturbed by late external noise (σ_{ext}). Intuitively, $F(e)$ captures how representational capacity is distributed efficiently, allocating greater precision to commonly experienced emotion intensities. The inverse transformation F^{-1} implements Bayesian decoding, reintroducing prior expectations and pulling uncertain estimates toward the prior peak. Prior mean and variance were set per participant and emotion from their empirical ratings, and inference combined these priors with noisy internal signals.

Fitted separately for each participant and emotion, the model reproduced rating variability across participants and emotions in Study 1 (Pearson's r=0.77, RMSE = 0.025, MAE = 0.019; Fig. 1d). Using the posterior emotion-intensity estimates in a choice rule, with the inputs fixed from the rating-model fits and only the internal noise (σ^2_{int}) and external noise (σ^2_{ext}) treated as free parameters, the model accurately predicted participants' trial-by-trial choices (average negative log-likelihood per trial = 0.61, AUROC = 0.78; Fig. 1e), their overall

choice consistency (average negative log-likelihood per trial = 0.89, AUROC = 0.97), and the empirical dependencies of choice consistency on rating difference (mean r across participants = 0.98, RMSE = 0.02; Fig. 1f) and summed variability (mean r across participants = 0.94, RMSE = 0.02; Fig. 1g). Participants' mean fitted rating variability further predicted their slope of rating difference on choice consistency in a mixed-effects logistic regression ($\beta_{\text{robust}} = -18.70 \pm 2.62$, $p < 0.001$; Fig. 1j), indicating reduced sensitivity when internal estimates are noisier.

All results replicated in the independent cohort, where the model again captured rating variability (r = 0.73), predicted trial-by-trial choices (average negative log-likelihood per trial = 0.50, AUROC = 0.85), predicted choice consistency (average negative log-likelihood per trial = 1.46, AUROC = 0.94), and reproduced the observed relationships with rating difference (mean r across participants = 0.93, RMSE = 0.03) and summed variability (mean r across participants = 0.87, RMSE = 0.04; Supplementary Fig. S10b-h).

Together, these results show that a single model combining efficient encoding with Bayesian decoding accounts for both the direction of bias and the precision observed in ratings and choices. This framework provides a mechanistic account of how uncertainty shapes emotion judgements and bridges perceptual and affective computation.

2.6 Mechanisms of emotion judgements and anxiety symptoms

We next asked whether individual differences in anxiety symptoms were related to the inferential mechanisms of emotion judgements. Given the central role of priors in our model, we hypothesised that participants with higher GAD-7 scores would show stronger and narrower priors, steeper Bayesian attraction slopes, and reduced confidence in their ratings and choices.

Participants in the replication study completed the GAD-7 questionnaire [33] (N=120; Supplementary Fig. S10a). At the subject level, higher GAD-7 scores predicted higher mean anxiety ratings ($\beta = 1.25$, 95% CI $[0.23, 2.27]$, $p = 0.017$; Fig. 6a), but had no reliable effect on rating variability ($\beta = 0.79$, 95% CI $[-1.77, 3.35]$, $p = 0.54$; Fig. 6b). GAD-7 scores were also not associated with the strength of Bayesian attraction slopes ($\beta = 0.0005$, 95% CI $[-0.007, 0.008]$, $p = 0.90$; Fig. 6c). Thus, although participants with higher anxiety symptoms gave higher average anxiety ratings, we found no evidence that they showed stronger and narrower priors.

We next tested these hypotheses in a larger study focused on anxiety judgements (Study 5, N=229). Participants reporting anxiety symptoms were recruited to complete a version of the task similar to the replication study, but with 95 video clips eliciting anxiety and no emotion discrimination phase (Supplementary Fig. S15a). The key Bayesian inference and efficient-coding effects were again observed in this cohort (Supplementary Fig. S15b-e), allowing us to test whether anxiety symptoms modulated these signatures.

In this sample, higher GAD-7 scores were not significantly associated with higher mean anxiety ratings ($\beta = 0.02$, 95% CI $[-0.003, 0.04]$, $p = 0.08$; Fig. 6f). Contrary to our prediction, higher GAD-7 scores were associated with wider, rather than narrower, priors ($\beta = 0.01$, 95% CI $[0.001, 0.019]$, $p = 0.038$; Fig. 6g). However, this difference in prior width did not translate into stronger Bayesian attraction: GAD-7 scores were not significantly related to attraction slopes ($\beta = 0.004$, 95% CI $[-0.001, 0.009]$, $p = 0.16$; Fig. 6h).

Finally, we tested whether anxiety symptoms were associated with reduced confidence. In the replication sample, GAD-7 scores showed numerically negative relationships with confidence in both emotion ratings and emotion-based choices, but neither effect reached significance (ratings: $\beta = -0.02$, 95% CI $[-0.04, 0.01]$, $p = 0.14$; choices: $\beta = -0.02$, 95% CI $[-0.04, 0.01]$, $p = 0.22$; Fig. 6d,e). In the larger anxiety sample, GAD-7 scores were also not associated with confidence in emotion ratings ($\beta = 0.004$, 95% CI $[-0.016, 0.023]$, $p = 0.71$; Fig. 6i).

Overall, anxiety symptoms showed little systematic association with the inferential mechanisms underlying emotion judgements. Although higher GAD-7 scores were associated with wider priors in the larger sample, this did not translate into altered prior attraction or reduced confidence. Thus, the core Bayesian and efficient-coding signatures of emotion judgements appeared largely preserved across variation in anxiety symptoms.

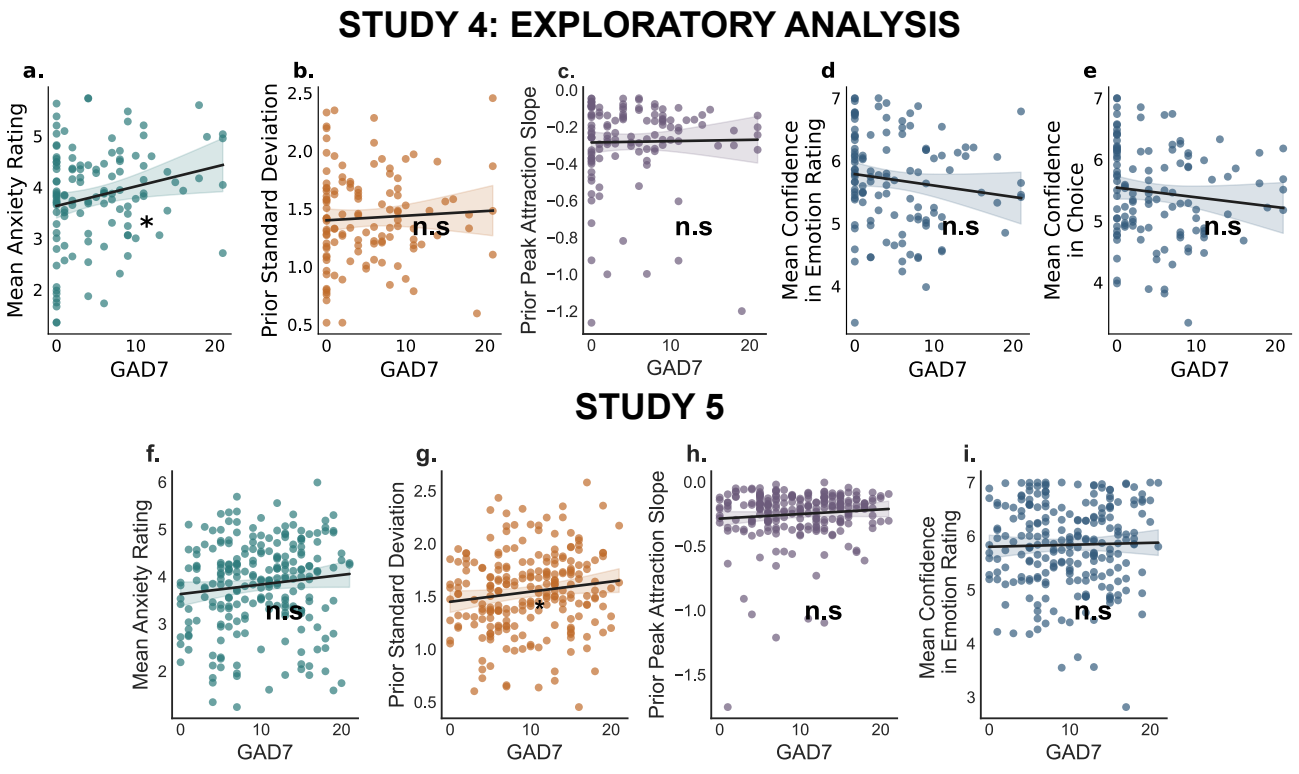

Fig. 6 Mechanisms of emotion judgements and anxiety symptoms in Study 4 (N=120) and Study 5 (N=229). **a, f. Emotion ratings and GAD-7.** Regression between mean anxiety rating per participant and GAD-7 score in Study 4 (**a**) and Study 5 (**f**). **b, g. Prior width and GAD-7.** Regression between participant-level prior standard deviation and GAD-7 score in Study 4 (**b**) and Study 5 (**g**). **c, h. Bayesian attraction to prior peak and GAD-7.** Regression between participant-level Bayesian attraction slopes, estimated from the relationship between short-long rating differences and long-duration ratings, and GAD-7 score in Study 4 (**c**) and Study 5 (**h**). **d, i. Confidence in emotion ratings and GAD-7.** Regression between mean confidence in emotion ratings per participant and GAD-7 score in Study 4 (**d**) and Study 5 (**i**). **e. Confidence in choices and GAD-7.** Regression between mean confidence in emotion discrimination choices per participant and GAD-7 score in Study 4. Shaded regions show 95% confidence intervals. Asterisks indicate statistical significance (* $p < 0.05$); n.s., not significant.

3 Methods

3.1 Participants

The main study tested 150 healthy young volunteers (mean age 38 years; 71 females) recruited via Prolific and randomly assigned to three experiments: Experiment 1 (n=57, 31 females), Experiment 2 (n=46 new participants, replication of Experiment 1; 22 females), and Experiment 3 (n=47 new participants, 18 females). Sample size was determined based on previous similar studies and pilot results. The replication study tested 120 healthy young volunteers (46 females) and the sample size was determined by a power analysis of the difference in consistency between choices near and far from prior peak observed in study 1. Participants had no neurological or psychological disorders and did not take medication that could affect participation. Finally, we conducted a study (Experiment 5) in a group of 229 participants (155 females) self-reporting anxiety symptoms on Prolific. All participants provided written informed consent and were compensated monetarily. All procedures were approved by the University College London Research Ethics Committee (approval ID 1896) and were conducted in accordance with the Declaration of Helsinki.

3.2 Task structure

Experiment 1 consisted of four main phases: (1) viewing phase, (2) rating phase 1, (3) rating phase 2, and (4) the emotion discrimination phase. Experiments 2 and 3 included only the viewing phase and two rating phases. Experiment 4, the replication study, included the viewing phase, the two rating phases, the emotion discrimination phase and two self-report questionnaires: GAD-7 and PHQ-9. Experiment 5 included the viewing phase, the two rating phases and the GAD-7 questionnaire.

In the viewing phase of all experiments, participants watched validated emotional video clips eliciting emotions across a wide intensity range [13].

3.3 Exclusion criteria

Across all studies, participants were excluded if they failed to meet basic attention or engagement requirements. Specifically, participants were excluded if they:

- Missed three or more attention-check questions during the viewing phase, or
- Failed to provide responses on ten or more trials in total, or missed three consecutive trials within a single phase of the study (rating or emotion discrimination).

All analyses were performed on the remaining participants who passed these criteria. A total of 91 participants were excluded in Study 1, 44 in Study 2, 38 in Study 3, 75 in the replication study, and 76 in the group of participants reporting anxiety symptoms. The final analysed samples were therefore: Study 1 = 57, Study 2 = 46, Study 3 = 47 (joy: N=42, anxiety: N=41, both: N=36), Study 4 = 120, Study 5 = 229.

3.4 Study 1

3.4.1 Viewing Phase

Participants watched videos that elicited five target emotions: joy, romance, anxiety, disgust, and sadness. A total of 142 videos were selected: joy (n=30), romance (n=27), anxiety (n=30), disgust (n=28), and sadness (n=27). Videos covered a range of intensities, selected across five intensity bins based on ratings from [13]. The video clips were presented in a pseudo-random order, structured into balanced blocks of four, ensuring uniform emotion and intensity distributions. Attention checks (yes/no questions about video content) appeared randomly once per block (n=16 correct "yes," n=19 correct "no").

3.4.2 Emotion Rating Phases 1 and 2

Participants completed two rating phases, rating video snapshots of the video clips previously watched during the viewing phase, on a slider scale. Importantly, because the participants saw all video clips before the ratings, they could effectively use the full range of the rating scale. Participants indicated "how strongly this video made them feel emotion," with each snapshot associated with a single target emotion. The slider scale was continuous, with both the numbers (1 to 7) and corresponding labels displayed: "not at all," "barely," "a little," "somewhat," "strongly," "very strongly," and "extremely". Participants were informed that the rightmost endpoint represented video clips eliciting extreme emotional intensity, while the leftmost endpoint represented video clips eliciting no emotional intensity. The snapshots were presented in a random order with randomised slider starting positions to minimize anchoring effects. Each trial began with a fixation cross presented for 500ms. This was followed by the emotion name, the video snapshot, a black screen with the word "think", and then the rating scale. Participants were informed that during the black screen, they should reflect on how strongly the video clip made them feel the emotion specified at the beginning of the trial. They were also instructed to provide their rating as fast as possible once the scale appeared on the screen.

Ratings were collected twice to assess variability in emotion intensity judgements. Rating phase 2 was identical to rating phase 1 and took place immediately after phase 1. The order of the video snapshots' presentation was randomized. Crucially, participants were not informed before the rating phase 1 that a second rating phase would take place, preventing them from intentionally memorising their initial ratings. Snapshots were presented either shortly (900 ms) or for a longer duration (2600 ms), with duration randomized in phase 1 and reversed in phase 2. The exposure times was pseudo-randomly selected for each video snapshot in the first round of ratings. This duration manipulation was applied during the black "think" screen preceding ratings. Crucially, participants were not informed about the exposure time manipulations.

3.4.3 Emotion Discrimination Phase

Immediately after the two rating phases, an algorithm selected a balanced set of decision trials divided into five emotion intensity rating difference levels on the rating scale (rating difference ∼5%, ∼10%, ∼15%, ∼20% and ∼50% of the length of the rating scale), as defined by the average rating across phases 1 and 2 provided by each participant. Emotion discrimination trials started with central presentation of a fixation cross for 500 ms. Immediately after this, two video snapshots were displayed simultaneously, one on the left and one on the right

field of the screen. The video snapshots were presented until response and participants had up to 3s to make a choice. Participants were asked to imagine they were actors tasked with portraying a specific emotion. On each trial, they were instructed to choose the video they felt would best set them in the mood for the emotion indicated at the beginning of the trial. To make their choice, participants used their mouse to select the video. Every four trials, participants were also asked to indicate their confidence in their choice. The confidence question was presented directly after the choice, randomly within each block of four trials and was answered using a continuous scale raging from 1 to 7, with both the numbers (1 and 7) and corresponding labels displayed: "not confident at all" (leftmost side), and "extremely confident" (rightmost side). A choice was defined as consistent if the selected snapshot had a higher mean rating from the previous rating phases. The trials were fully balanced across rating difference levels and the location of the consistent response option (left or right).

3.5 Study 2

Study 2 replicated Study 1 with two modifications:

- The black screen with "think" was removed to streamline the trial sequence, and the manipulation of exposure time was applied to the video snapshot itself.
- The emotion discrimination phase was removed, focusing on how exposure duration affects emotion judgements.

3.6 Study 3

To strengthen prior expectations, Study 3 presented emotional material in emotion-specific blocks rather than in pseudo-randomised order.

3.6.1 Viewing Phase

Participants watched 160 videos (80 joy, 80 anxiety), selected to bias exposure toward mid-intensity emotions (bins 1, 5: 7 videos; bins 2, 4: 17 videos; bin 3: 32 videos). Videos were grouped in emotion specific blocks (first anxiety, then joy) and presented in random intensity order within blocks. Attention checks were incorporated randomly within blocks (n=16 correct "yes," n=19 correct "no").

3.6.2 Emotion Rating Phases 1 and 2

Participants first watched and completed the two rating phases for anxiety-inducing videos, followed by joy videos, with the video snapshots presented in a random order within each emotion block. Each trial began with a fixation cross presented for 500 ms, followed by the video snapshot and the rating scale. Participants were instructed to provide their rating as fast as possible once the scale appeared on the screen.

Rating phase 2 was identical to rating phase 1 and took place immediately after phase 1. The order of the video snapshots' presentation was randomized. Each snapshot was presented either for 900 ms or 2600 ms, with durations randomized in phase 1 and reversed in phase 2, for each emotion. The exposure time was pseudo-randomly selected for each video snapshot in the first round of ratings.

Every four trials, with the confidence trial randomly positioned within each block of four, participants were asked to indicate how confident they were that the rating they had provided accurately reflected how they truly felt about the video. Confidence was reported on a continuous scale ranging from 1 to 7, anchored by the labels "not confident at all" (1) and "extremely confident" (7).

3.7 Study 4: Replication study

The replication study was pre-registered (https://doi.org/10.17605/OSF.IO/NMBSU) and followed the same procedure as Study 3, with three modifications: (i) only the anxiety block was administered; (ii) the emotion discrimination task from Study 1 was included; and (iii) participants additionally completed the GAD-7 and PHQ-9 questionnaires at the end of the task. Each questionnaire included an embedded attention-check item instructing participants to select a specific response. No participants failed these checks, and all were retained for analysis.

3.8 Study 5

Study 5 followed the same procedure as the replication study, with three modifications: (i) the number of videos eliciting anxiety was increased to 95; (ii) the emotion discrimination task was omitted; and (iii) participants

completed only the GAD-7 questionnaire at the end of the task. Similar to study 4, the questionnaire included an embedded attention-check item instructing participants to select a specific response; no participants failed this check, and all were retained for analysis.

3.9 Computational model of efficient representation of emotion judgements

3.9.1 Model

Inspired by the efficient-coding model of subjective value developed by Polanía et al. [4], we adapted the same general framework to emotion judgements, modelling them as arising from noisy encoding and Bayesian decoding of a latent true emotion intensity. The presentation of a video clip with an underlying true emotion intensity e_0 elicits an internal noisy neural response e_{enc}, from which the observer derives a subjective emotion intensity estimate e_{dec}.

At the encoding stage, a function $F(e)$ maps the emotion intensity space to a new space where Fisher information is uniform. $F(e)$ is the cumulative distribution function (CDF) of the prior distribution $p(e)$, defined as:

$$F(e) = \int_0^e p(\chi) d\chi \quad , \tag{1}$$

where $p(\chi)$ represents the prior distribution over emotion intensities. This transformation ensures efficient allocation of neural resources, assigning higher encoding precision to frequently encountered intensities. Encoding adds Gaussian noise to this transformed representation:

$$e_{\text{enc}} = F(e) + \epsilon_{\text{int}} \tag{2}$$

Where internal encoding noise $\epsilon_{\text{int}} \sim \mathcal{N}(0, \sigma^2_{\text{int}})$ remains constant across all intensity levels, capturing trial-to-trial variability in reported emotion intensity ratings.

At the decoding stage, the observer reconstructs the emotion intensity estimate from the encoded signal via the inverse transformation:

$$e_{\text{dec}} = F^{-1}(e_{\text{enc}}) \tag{3}$$

External noise is introduced to account for late noise in the decision stage (e.g., noise introduced during response selection). This results in a final reported estimate:

$$e_{\text{dec}} = F^{-1}(e_{\text{enc}}) + \epsilon_{\text{ext}} \tag{4}$$

where $\epsilon_{\text{ext}} \sim \mathcal{N}(0, \sigma^2_{\text{ext}})$ represents post-decoding noise, capturing downstream variability that is unrelated to the encoding process itself.

Considering both internal and external noise, the expected reported emotion intensity given a true stimulus intensity e_0 is:

$$E[\hat{e} \mid e_0] = E[F^{-1}(F(e_0) + \epsilon_{\text{int}}) + \epsilon_{\text{ext}}] \tag{5}$$

The likelihood of observing a given reported emotion intensity e_{dec} conditioned on a true stimulus value e_0 is given by:

$$p(e_{\text{dec}} \mid e_0) = \int de_{\text{enc}}\, p(e_{\text{dec}} \mid e_{\text{enc}}) \cdot p(e_{\text{enc}} \mid e_0) \tag{6}$$

where:

$$p(e_{\text{dec}} \mid e_{enc}) = \mathcal{N}(F^{-1}(e_{\text{enc}}), \sigma^2_{\text{ext}}) \tag{7}$$

$$p(e_{\text{enc}} \mid e_0) = \mathcal{N}(F(e_0), \sigma^2_{\text{int}}) \tag{8}$$

Since there is no closed-form solution, this integral was approximated numerically by sampling from the conditional distributions.

3.9.2 Model Implementation and Fitting

To simplify model estimation, we assumed a normal distribution prior over emotion intensities. Thus, the encoding function $F(e)$ corresponds to the cumulative distribution function (CDF) of a normal distribution:

$$F(e) = \Phi\left(\frac{e - \mu_{\text{prior}}}{\sigma_{\text{prior}}}\right) \tag{9}$$

where μ_{prior} and σ_{prior} define the prior mean and standard deviation, respectively.

Since there is no closed-form solution for the likelihood integral, we used a sampling approach to approximate the likelihood. The model was implemented in Turing.jl, a probabilistic programming library in Julia, enabling efficient Bayesian inference via Hamiltonian Monte Carlo (HMC).

Before model fitting, empirical ratings were normalised to the [0, 1] range to facilitate comparison with model predictions. For study 1, the model was fitted separately for each participant and emotion to emotion intensity ratings obtained in the two rating phases. A single internal noise parameter (σ^2_{int}) was shared across exposure times. Prior parameters (μ_{prior}) and (σ_{prior}) were derived directly from that participant's own rating distribution (phases 1 and 2 combined). The model then estimated the internal (σ^2_{int}) and external noise (σ^2_{ext}), and the latent true stimulus emotion intensity ratings (e_0), by maximising the likelihood of observed ratings, subject to the constraint that e_0 followed the specified prior distribution.

To quantify the model's accuracy in reproducing observed rating variability, we generated simulated ratings from 500 posterior samples of e_0, propagating these through the model to produce predicted ratings e_{dec} for each rating phase. The standard deviation of these simulated ratings provided an estimate of rating variability, which was compared to participants' empirical rating variability. To match the bounded rating scale used in the task, simulated e_{dec} values were rescaled to the [0, 1] interval using a logistic (sigmoid) transformation:

$$e_{\text{dec}}^{\text{bounded}} = \text{logistic}\left(\frac{e_{\text{dec}} - m}{s}\right), \tag{10}$$

where m and s set the midpoint and scaling based on the 2nd–98th percentile range of simulated values.

Given model-estimated stimulus emotion intensity ratings ($\hat{e}_1, \hat{e}_2$) and prior parameters ($\mu_{\text{prior}}, \sigma^2_{\text{prior}}$), the probability of choosing one alternative over another was computed as:

$$P(\hat{e}_1 > \hat{e}_2 \mid e_1, e_2) = \Phi\left(\frac{E[\hat{e}_1 \mid e_1] - E[\hat{e}_2 \mid e_2]}{\sqrt{\text{Var}[\hat{e}_1 \mid e_1] + \text{Var}[\hat{e}_2 \mid e_2] + 2\sigma^2_{\text{ext}}}}\right) \tag{11}$$

Where Φ is the CDF of the normal distribution.

The inputs to the choice model were fixed from the rating model fits, leaving only two free parameters: the internal noise (σ^2_{int}) and the external noise (σ^2_{ext}). The choice model was implemented in Turing.jl using a Bernoulli likelihood, where choices were modeled as binary outcomes.

Both rating and choice models were estimated using HMC with the No-U-Turn Sampler (NUTS), run for 10,000 iterations across three chains, with initial samples discarded as burn-in and tuning acceptance thresholds for efficient exploration of the posterior distribution. We employed uniform priors for both noise parameters (σ_{int},$\sigma_{\text{ext}} \sim \text{Uniform}(0,0.1)$).

The same procedure was applied in the replication study, yielding comparable model performance across cohorts.

3.10 Behavioural Analyses and Statistics

Emotion intensity rating variability was quantified as the standard deviation of ratings across phases 1 and 2 for each video. Choice consistency in the discrimination phase was modelled using mixed-effects logistic regressions with random intercepts and random slopes per subject and emotion. Separate models tested the effects of absolute rating difference and summed variability. Choice consistency was then modelled using logistic regressions per emotion with the following fixed predictors: (1) rating difference (absolute difference in mean ratings), (2) summed variability (sum of standard deviations of the two clips), and (3) summed ratings (sum of mean ratings). Model improvement when adding variability was assessed with likelihood-ratio tests, reported as $\Delta\chi^2$. To test whether variability systematically reduced sensitivity to rating differences, we fit a mixed-effects logistic regression with random intercepts and random slopes for rating difference per participant and emotion. Random slopes were then regressed against log-transformed variability using robust regression.

Reaction times in the emotion discrimination task were analysed as an additional behavioural signature of perceptual decision-making. Reaction times were log-transformed before analysis. For each emotion category, we fit linear mixed-effects models predicting log reaction time from choice consistency, and separately from the absolute rating difference between the two choice options, with random intercepts for participant.

Prior-driven bias was computed as the difference between short- and long-exposure ratings of the same video. Piecewise linear regressions of bias on long-exposure ratings were run separately above, below, and near the prior peak. For these analyses, ratings were standardized per participant and emotion, using the prior mode and standard deviation. The prior peak was defined as the mode of the empirical rating distribution (combined from rating phases 1 and 2), estimated with a boundary-corrected kernel density estimator (reflection at the [1,7] bounds). The prior standard deviation was defined as the empirical standard deviation of the same ratings. At the participant level, attraction slopes (from regressions of bias on mode-centred long-exposure ratings) were regressed on prior width.

As a control analysis, we tested whether the amount of information available in the snapshot itself modulated prior-driven bias. A subset of snapshots was manually classified as informative, that is, as containing sufficient visual information to identify the source video clip directly from the snapshot, whereas the remaining snapshots were classified as less informative. We then fit a linear mixed-effects model predicting rating bias from mode-centred long-exposure ratings, snapshot informativeness, and their interaction, with random intercepts for participant.

To test whether choice consistency varied with distance from the prior peak, we computed the midpoint rating of each choice pair and measured its absolute distance from the participant-specific prior mode, scaled by the participant-specific prior standard deviation. Analyses of discriminability were restricted to small rating differences (<0.6 units; ∼10% of the scale), as these corresponded to the most difficult choices. Within this subset, choice pairs were classified as near if their midpoint rating lay within 0.3 units of the prior mode, and far if the midpoint was more than 1.9 units away. These thresholds were based on the observed distribution of distances from the prior peak, chosen to clearly separate stimuli adjacent to versus distant from the peak while maintaining balanced trial counts. For visualisation and complementary analyses, trials were grouped into near, far, and large rating-difference categories. Choice consistency between near and far trials was compared using Mann-Whitney U tests at the trial level and paired t-tests on within-subject differences in mean consistency. We also fit logistic mixed-effects models predicting choice consistency from the continuous distance to the prior peak and the absolute rating difference between the two options, with random intercepts and random slopes for prior distance by participant.

Confidence ratings were analysed using mixed-effects linear models, with fixed effects of rating difference and summed variability (for choices) or rating variability and mean rating (for ratings), and random intercepts per subject.

For questionnaire analyses, we regressed mean ratings, variability, and slopes of attraction against GAD-7 scores at the subject level (OLS regressions). Confidence was analysed as above but with GAD-7 as predictors in mixed-effects models.

The computational model combined efficient coding with Bayesian decoding (Methods, Section 3.9.1). Priors were set from each participant's empirical rating distribution (empirical mean and standard deviation). Model fits were evaluated using Pearson's r, RMSE, and MAE for variability, and average negative log-likelihood (NLL) per trial and area under the receiver-operating characteristic (AUROC) for choices.

For Study 5, the same Bayesian inference and efficient-coding analyses were applied to the anxiety-only sample. Because the task did not include an emotion discrimination phase, analyses were restricted to ratings, confidence in ratings, prior width, and Bayesian attraction slopes. Associations with anxiety symptoms were tested by regressing mean anxiety ratings, prior standard deviation, attraction slopes and mean confidence ratings against GAD-7 scores at the participant level.

Mixed-effects models were implemented using either `pymer4`, which interfaces with `lme4` in R, or `statsmodels`, depending on the analysis. Ordinary least-squares regressions, generalized linear models and robust regressions were implemented using `statsmodels`. Non-parametric tests, paired t-tests and correlation analyses were implemented using `scipy.stats`. Statistical significance was set at $p < 0.05$ (two-sided), with thresholds at $p < 0.01$ and $p < 0.001$.

The same preprocessing and analysis procedures were applied to the independent replication study where the relevant task components were present. In the replication study, the corresponding near- and far thresholds were chosen analogously from the empirical distribution of distances to the prior peak to preserve separation between adjacent and distant trials while maintaining sufficient trial counts.

3.11 Data availability

The data is available (HERE).

3.12 Code availability

The code is available (HERE).

4 Discussion

Emotion judgements appear to behave like perceptual decisions, relying on probabilistic inference under representational constraints. By combining repeated ratings, a two-alternative forced-choice task, and a generative model grounded in efficient coding and Bayesian decoding, we show that variability in emotion-intensity ratings is not mere noise but reflects principled uncertainty shaped by prior experience. In this sense, emotion reports appear not as direct readouts of stable internal states, but as estimates sensitive to statistical uncertainty.

First, variability was structured rather than random, and predicted participants' emotion-based choices. Trial-level fluctuations in emotion ratings captured meaningful computational properties of the judgement process rather than measurement error [34, 35]. Repeated emotion ratings contained information about the latent reliability of the estimate giving rise to the judgement, much as variability in perceptual reports reflects uncertainty rather than merely poor responding. This has practical consequences: averaging away within-person variability may discard signal that is computationally informative.

Second, participants' ratings were biased toward empirically learned priors, especially under high uncertainty, consistent with Bayesian attraction. These observations provide a computational mechanism for why context and prior experience should shape emotion reports. This extends previous evidence that emotion inferences about others integrate prior expectations with ambiguous cues [36–38], and connects naturally to appraisal and constructed emotion theories, which similarly describe emotions as context-dependent inferences combining affective and conceptual information [1, 39–41]. Additionally, this opens a bridge to broader affective schemas, mood, or long-term learning: task-induced priors are local and experimentally tractable, but they may be one instantiation of more general affective expectations.

Third, efficient-coding predicted, and experiments confirmed, that representational precision is allocated according to the frequency of experienced emotion intensities. Participants showed two hallmark signatures of this process: repulsive bias near the prior peak and enhanced discriminability at frequently experienced intensities. These signatures are not predicted by generic prior bias alone, but point specifically to non-uniform coding precision, produced only when Fisher information scales with prior density. Emotion judgements may therefore be encoded in a resource-efficient manner, with precision scaled according to the statistical structure of past emotion experience. This extends efficient-coding principles beyond traditional sensory and value-based domains [3, 4]. Consistent with this frequency-dependent account, Goel et al. (2024) [38] showed that observers weight facial cues more strongly for emotion categories they encounter most often in everyday life. More broadly, it implies that emotional precision should be experience-dependent: individuals with different affective histories may become most precise in different regions of emotion space.

Fourth, confidence in emotion ratings decreased reliably as repeated ratings became more variable, consistent with the broader view that confidence tracks the reliability or uncertainty of an internal estimate [42–44]. Recent work further suggests that metacognition of one's own emotional states can be quantified, extending metacognitive analysis beyond standard perceptual decisions to subjective and affective judgements [45, 46]. Confidence in choices was more robustly related to rating difference, consistent with perceptual decision-making studies showing that confidence in discrete decisions scales with evidence strength and decision reliability [43, 47]. Together, these findings suggest that confidence in emotion judgements is sensitive to subjective uncertainty. Future work should test this more directly using formal models of confidence and metacognitive sensitivity.

We also asked whether these inferential signatures vary systematically with anxiety symptoms. Altered affective processing in depression and anxiety makes this plausible, as these conditions have been linked to shifts in the evaluation of emotional stimuli and to more negative affective interpretations [48, 49]. Related work further suggests that anxiety and depression can be associated with lower confidence or underconfidence in perceptual and memory judgements [50, 51]. In our data, however, anxiety symptoms showed limited associations with the computational signatures identified here. Taken together, these findings suggest that the core Bayesian and efficient-coding features of emotion judgement may be relatively preserved across variation in anxiety symptoms, although larger clinically enriched samples will be needed to test whether more specific symptom dimensions modulate them.

These findings have important implications for affective measurement. Standard self-report instruments often assume that emotion ratings provide direct access to stable internal states. Yet computational accounts linking subjective emotion reports to underlying cognitive and neural processes remain relatively limited, making it difficult to distinguish measurement noise from principled uncertainty in emotional self-report [16]. Our data instead suggest that reported intensities emerge from an inferential process shaped by learned statistical regularities and internal noise. Variability and bias in self-report should therefore not automatically be treated as error, but may instead reflect meaningful cognitive operations, with implications for both basic research and clinical assessment.

Several limitations remain. Our priors were induced by a restricted set of task stimuli and therefore reflect local, experimentally induced expectations rather than broader, lifelong affective schemas or mood states. In addition, although our model captures the main empirical signatures, it relies on relatively simple prior and noise assumptions. Future work should test whether these conclusions generalise across richer stimulus statistics, developmental and cultural contexts, and more flexible computational models.

In sum, these findings bridge affective science with computational theories of perception and decision-making, suggesting that emotion judgements arise from probabilistic inferences shaped by prior expectations and implemented through capacity-limited encoding processes. Treating variability not as noise but as signal offers a more mechanistic account of how people access and report their own emotional experiences.

5 Acknowledgements

We thank J. Chen for assistance with task implementation and for helpful guidance on `jsPsych` and Y. Abir for helpful guidance on `JULIA`. We are grateful to T. Sharot for helpful discussions and for input on task development. We also thank members of the Applied Computational Psychiatry Lab for helpful discussions. Finally, we thank all participants for taking part in the study.

6 Funding

This work was funded by Biotechnology and Biological Sciences Research Council (BB/T008709/1) and Wellcome Trust (221826/Z/20/Z). JRS was supported by the Biotechnology and Biological Sciences Research Council (BB/T008709/1). QJMH has received research grant funding from Carigest S.A., Koa Health, NIHR and Wellcome Trust. We acknowledge support by the NIHR UCLH BRC.

7 Author information

Authors and Affiliations

Applied Computational Psychiatry Lab, Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology and Mental Health Neuroscience Department, Division of Psychiatry, University College London, London, UK

Jade R. Serfaty & Quentin J. M. Huys

Contributions

J.R.S. and Q.J.M.H. conceived the study and developed the computational modelling framework. J.R.S. designed the experiments, collected the data and performed the analyses. Q.J.M.H. supervised experimental design, data collection and analysis. Both contributed to the conceptual framing and interpretation of the results. J.R.S. wrote the original draft of the manuscript. Q.J.M.H. reviewed and edited the manuscript.

Corresponding author

Correspondence to Jade R. Serfaty and Quentin J. M. Huys.

8 Ethics declarations

Competing interests

QJMH was employed by University College London during this work. QJMH has obtained fees and options for consultancies for Aya Technologies and Alto Neuroscience.

9 Extended data figures

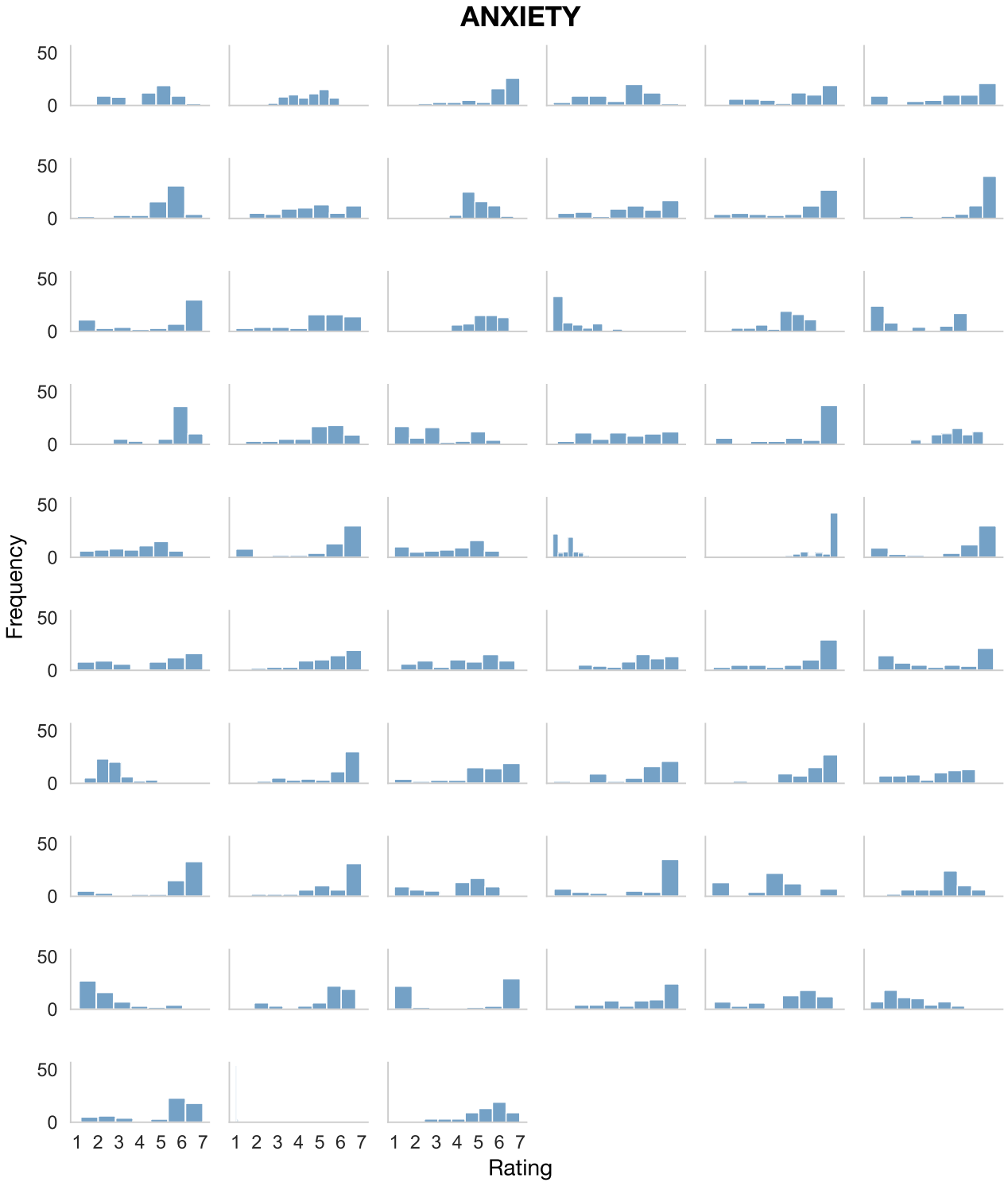

Fig. S1 Distribution of anxiety ratings across participants in Study 1 (N=57). Each small panel corresponds to one participant and shows the frequency of ratings across anxiety videos on the 1-7 rating scale.

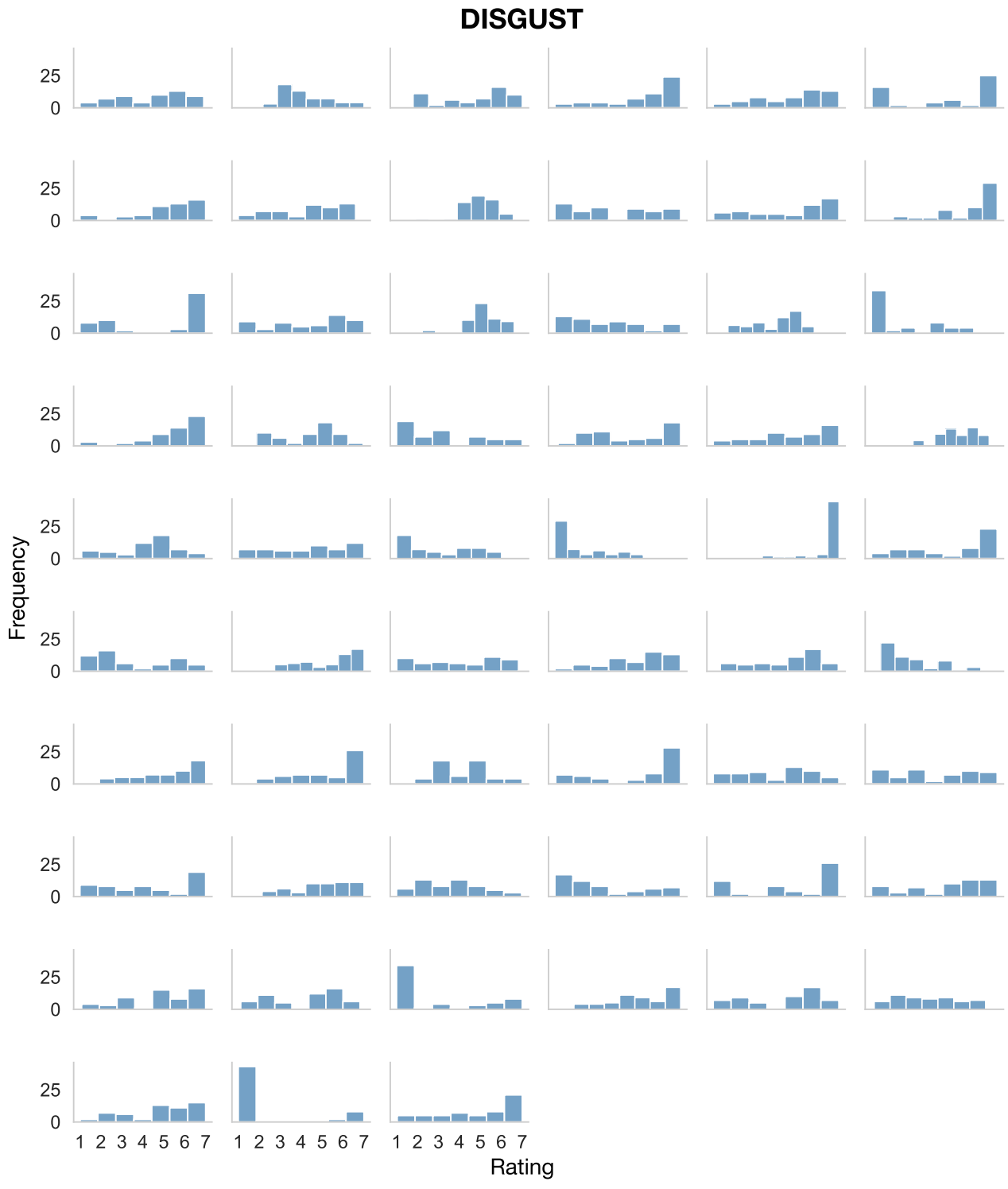

Fig. S2 Distribution of disgust ratings across participants in Study 1 (N=57). Each small panel corresponds to one participant and shows the frequency of ratings across disgust videos on the 1-7 rating scale.

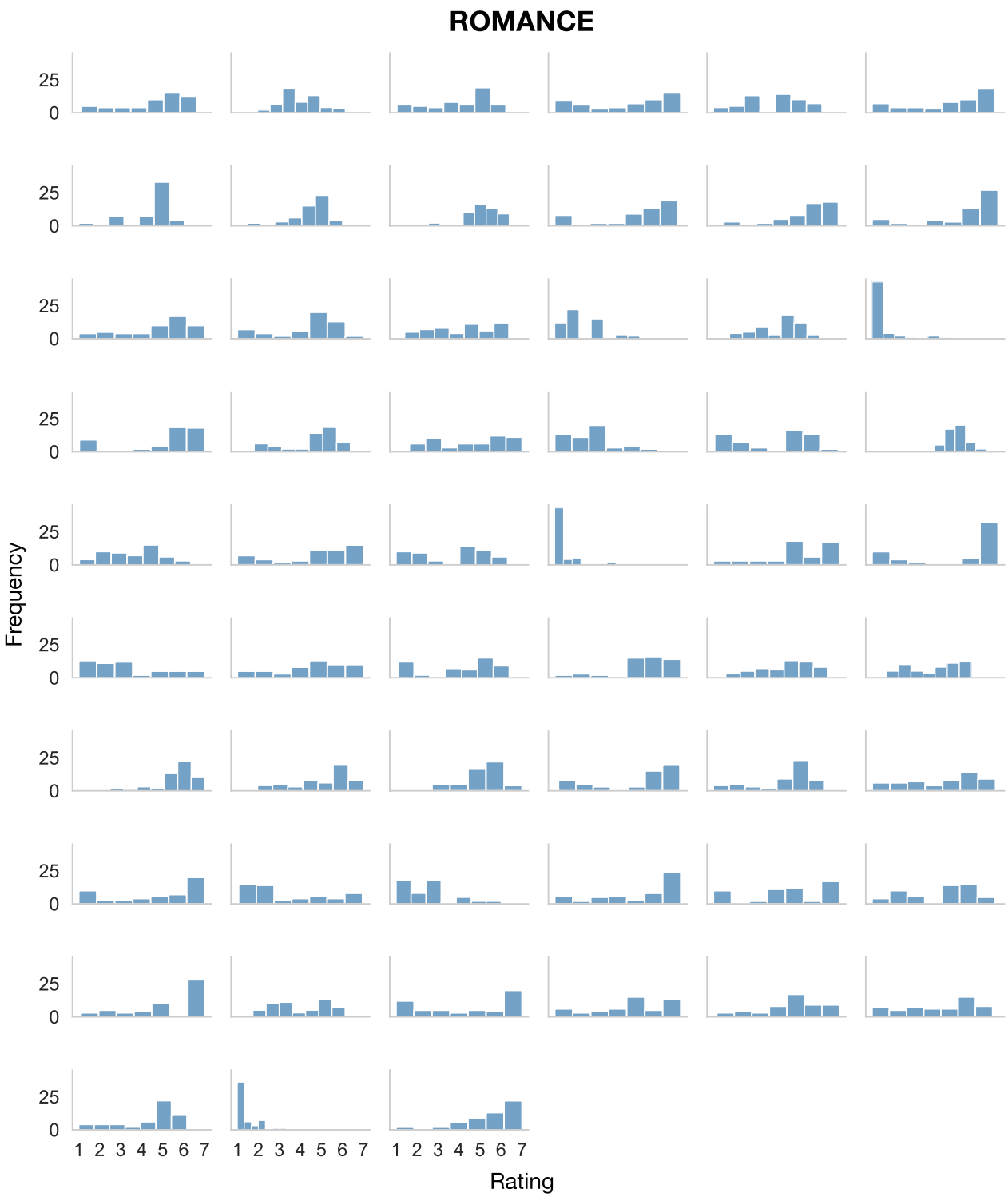

Fig. S3 Distribution of romance ratings across participants in Study 1 (N=57). Each small panel corresponds to one participant and shows the frequency of ratings across romance videos on the 1-7 rating scale.

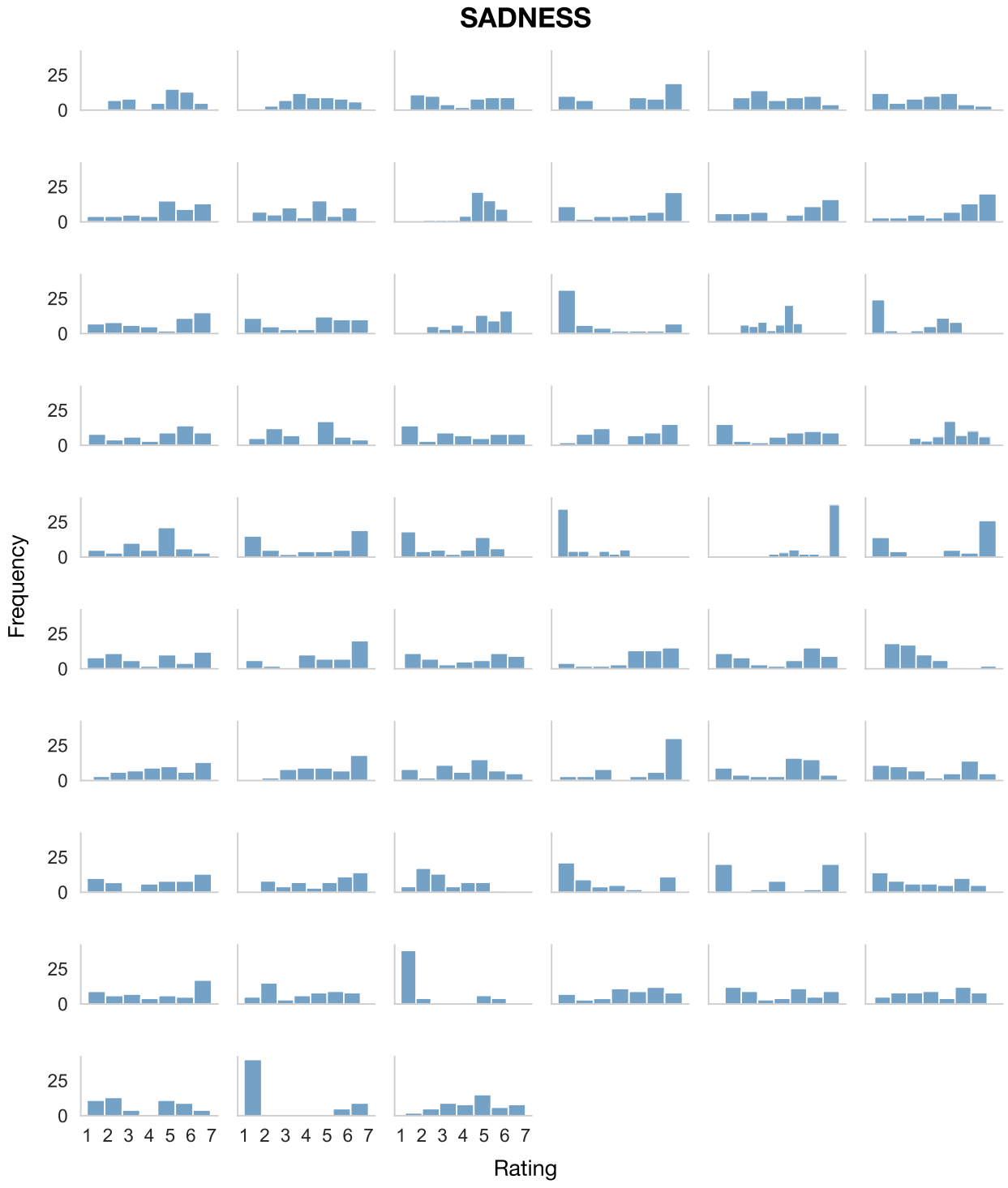

Fig. S4 Distribution of sadness ratings across participants in Study 1 (N=57). Each small panel corresponds to one participant and shows the frequency of ratings across sadness videos on the 1-7 rating scale.

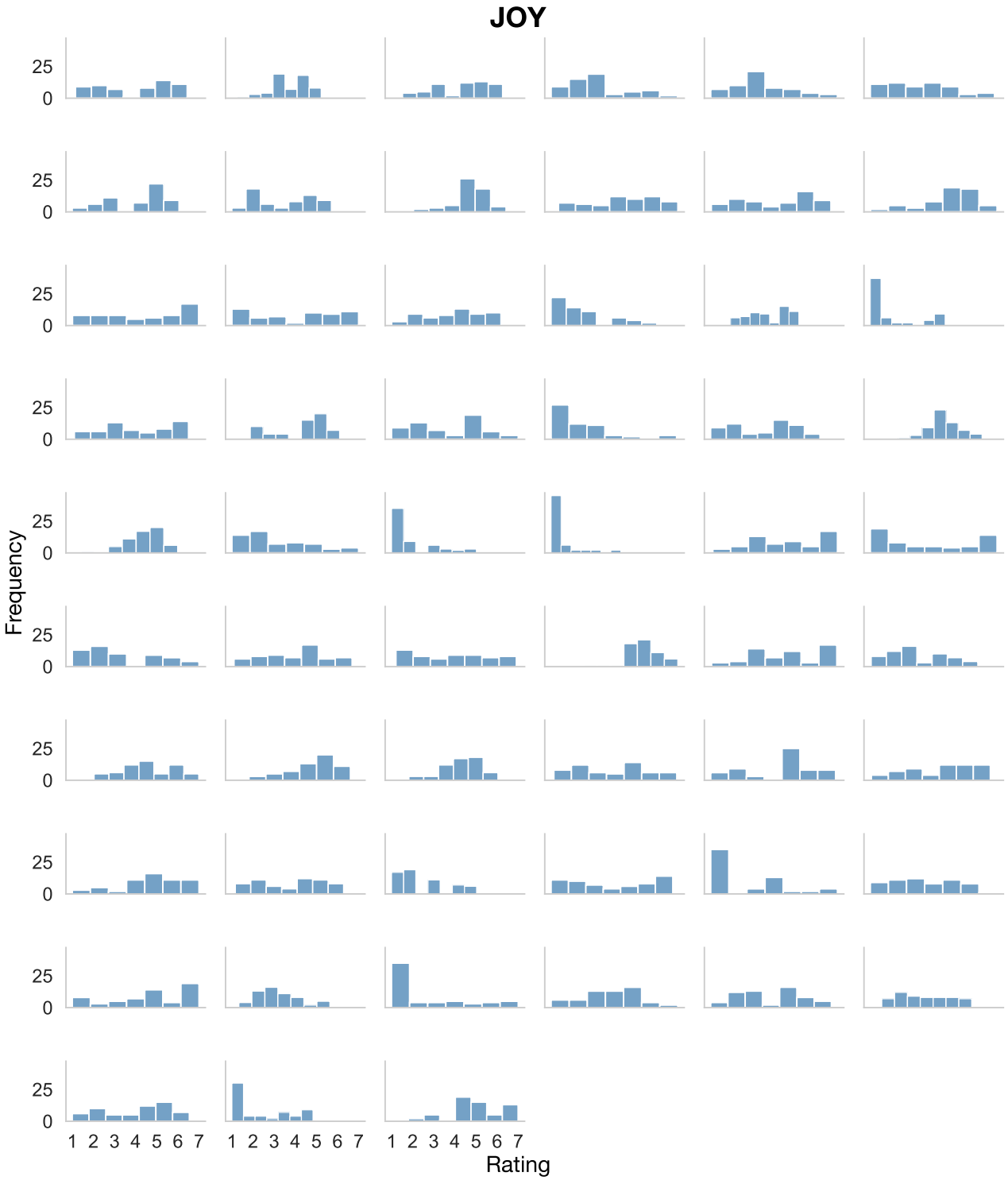

Fig. S5 Distribution of joy ratings across participants in Study 1 (N=57). Each small panel corresponds to one participant and shows the frequency of ratings across joy videos on the 1-7 rating scale.

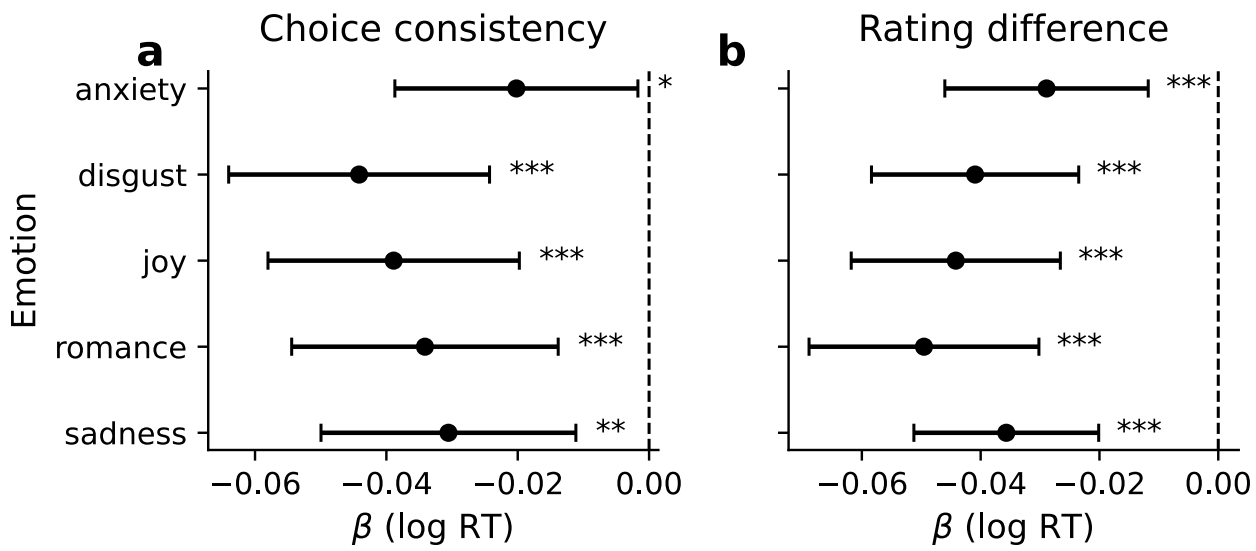

Fig. S6 Reaction times in the emotion discrimination task (Study 1; N=57). **a.** Fixed-effect coefficients from linear mixed-effects models predicting log reaction time from choice consistency, fit separately for each emotion with random intercepts for participant. Negative coefficients indicate that consistent choices were made faster than inconsistent choices. **b.** Fixed-effect coefficients from linear mixed-effects models predicting log reaction time from the absolute rating difference between the two choice options, fit separately for each emotion with random intercepts for participant. Negative coefficients indicate that choices were faster when the two options differed more strongly in their mean emotion ratings. Points show fixed-effect estimates and horizontal bars show 95% confidence intervals. Asterisks indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

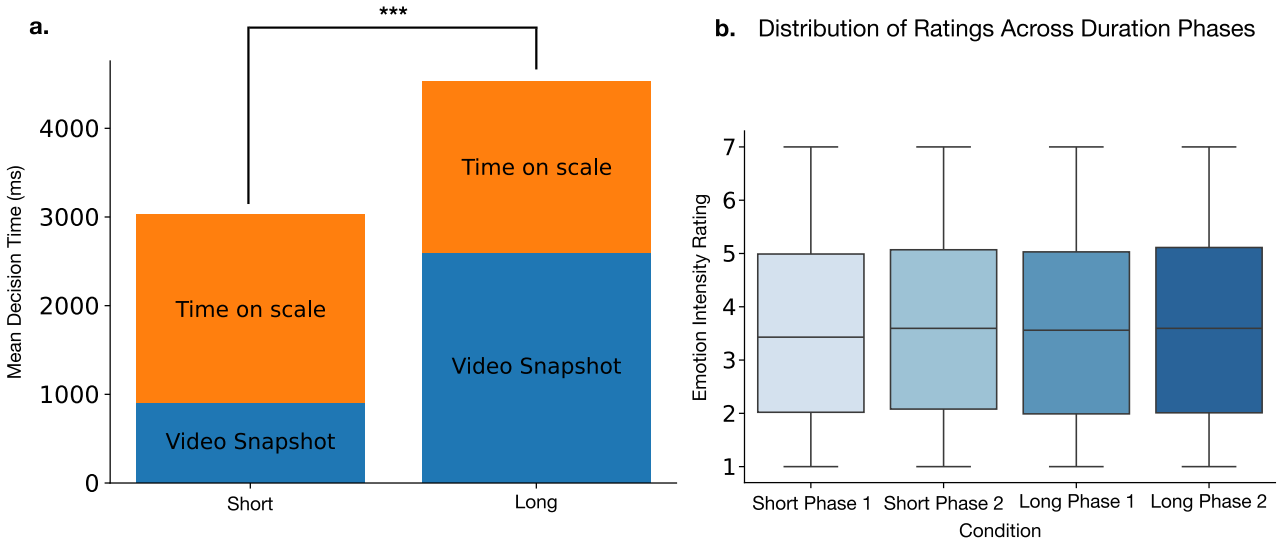

Fig. S7 Control analyses for exposure duration and rating phase in Study 3 (N=47). **a.** Mean total rating-trial time, defined as video snapshot duration plus time spent on the rating scale, in the short- and long-duration conditions. As expected, long-duration trials had significantly longer total trial times than short-duration trials (***$p < 0.001$). **b.** Distribution of emotion intensity ratings across rating phases and duration conditions. Boxplots show similar rating distributions across Phase 1 and Phase 2 for both short- and long-duration trials, indicating no systematic phase-related shift in ratings. Boxplot centre lines show medians; boxes show 25th-75th percentiles; whiskers extend to 1.5× IQR.

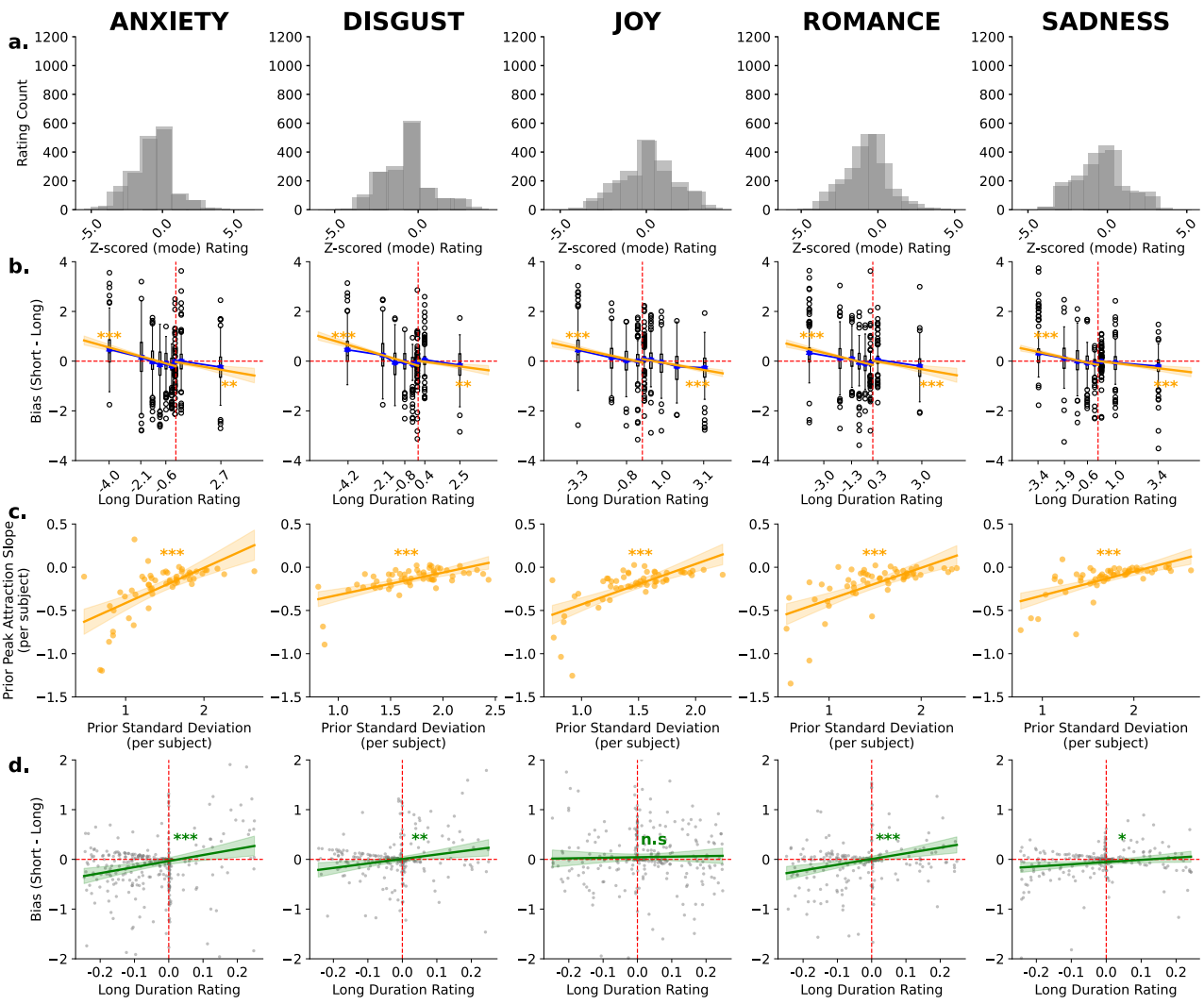

Fig. S8 Bayesian attraction and efficient-coding diagnostics in Study 1 (N=57). **a.** Aggregated participant-centred rating distributions for each emotion category. Ratings were z-scored and aligned to each participant's modal rating. **b.** Bias in emotion intensity judgements, quantified as the short-long rating difference, plotted against long-duration ratings for each emotion. Red dashed lines indicate zero bias and the participant-specific prior peak. Blue points show binned means; orange lines show piecewise linear fits ± 95% CI. **c.** Relationship between participant-level prior standard deviation and Bayesian attraction slope. Each point represents one participant; orange lines show OLS fits ± 95% CI. **d.** Bias close to the prior peak, plotted against long-duration ratings. Green lines show linear fits ± 95% CI testing for repulsive bias near the prior peak. Asterisks indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$); n.s., not significant.

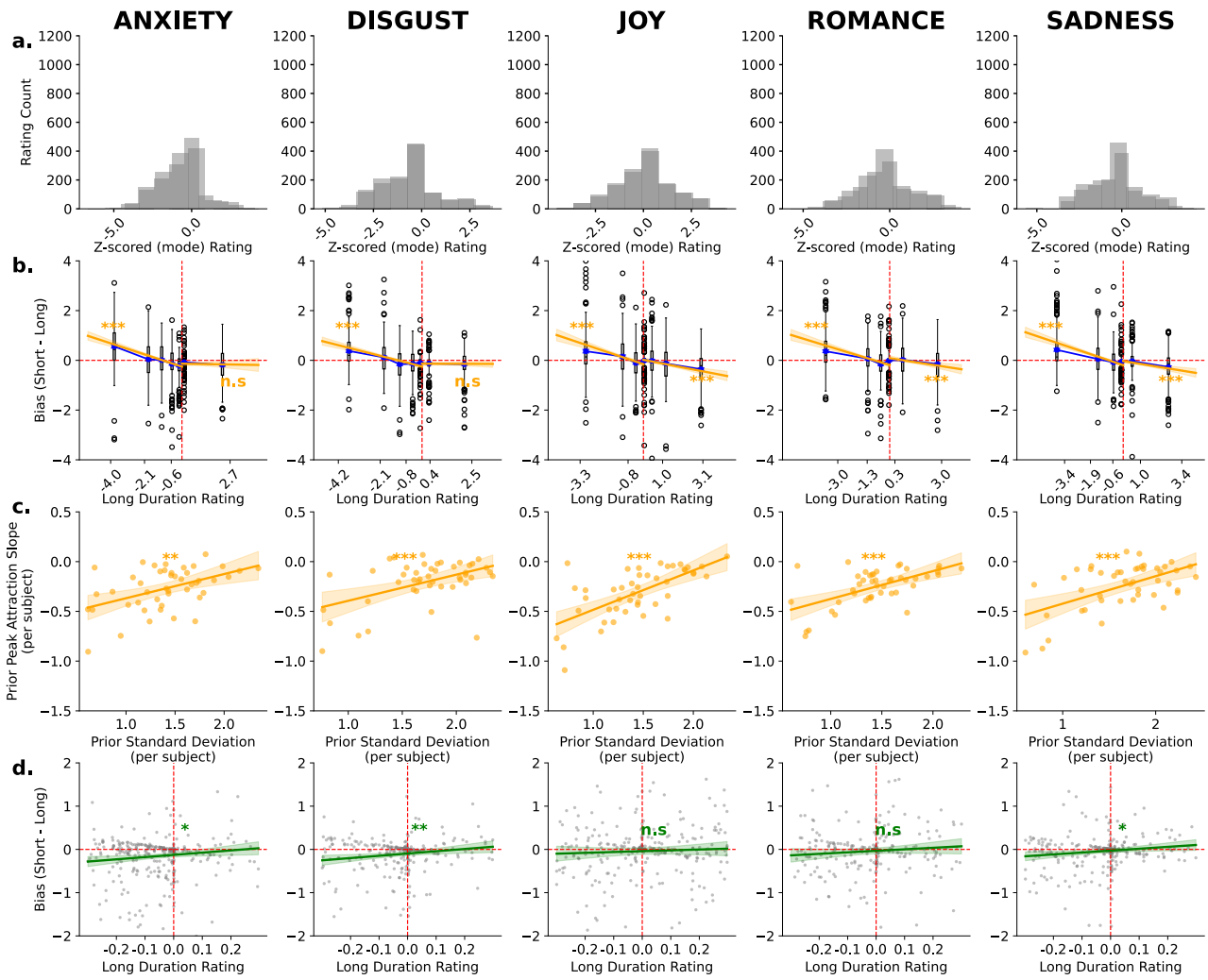

Fig. S9 Bayesian attraction and efficient-coding diagnostics in Study 2 (N=46). **a.** Aggregated participant-centred rating distributions for each emotion category. Ratings were z-scored and aligned to each participant's modal rating. **b.** Bias in emotion intensity judgements, quantified as the short-long rating difference, plotted against long-duration ratings for each emotion. Red dashed lines indicate zero bias and the participant-specific prior peak. Blue points show binned means; orange lines show piecewise linear fits ± 95% CI. **c.** Relationship between participant-level prior standard deviation and Bayesian attraction slope. Each point represents one participant; orange lines show OLS fits ± 95% CI. **d.** Bias close to the prior peak, plotted against long-duration ratings. Green lines show linear fits ± 95% CI testing for repulsive bias near the prior peak. Asterisks indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$); n.s., not significant.

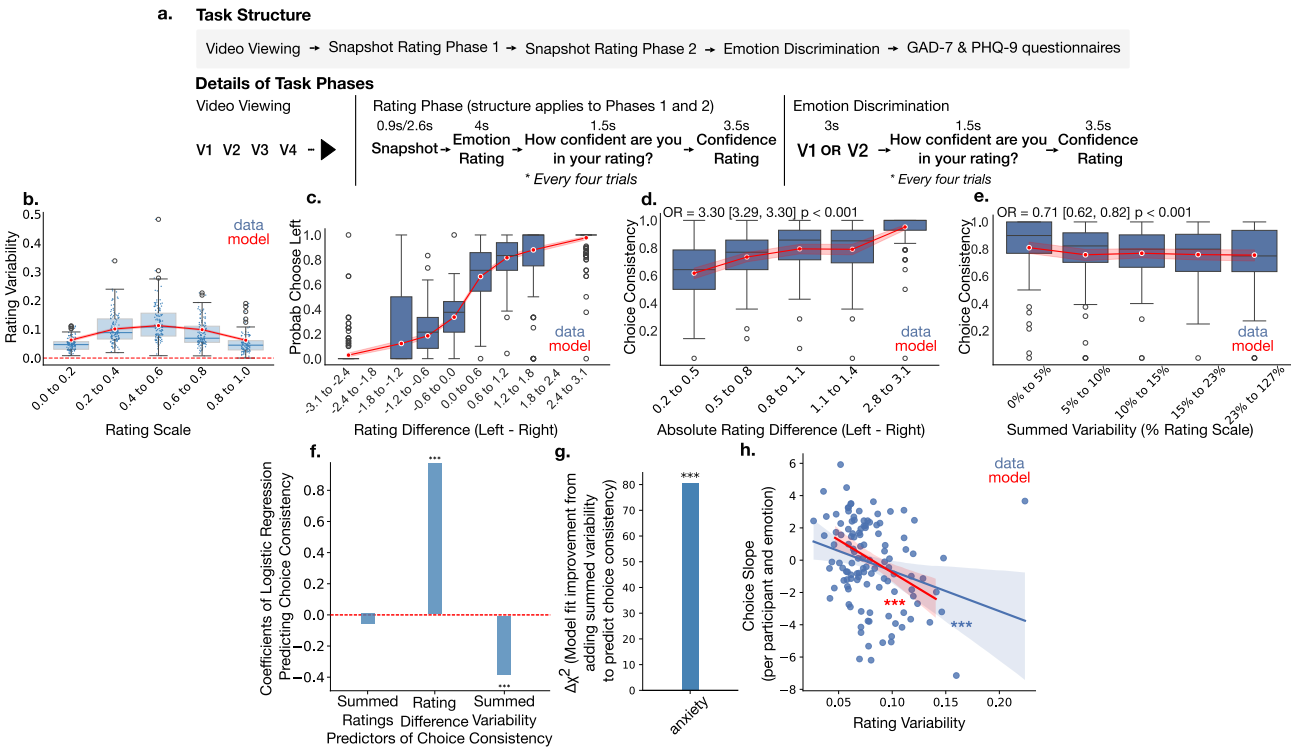

Fig. S10 Replication of rating variability effects in Study 4 (N=120). **a. Task structure.** Participants completed video viewing, two snapshot rating phases, the emotion discrimination task, and GAD-7 and PHQ-9 questionnaires. In each rating phase, participants rated anxiety elicited by each previously viewed video snapshot and reported confidence every four trials. In the emotion discrimination task, participants chose which of two previously rated videos would better induce anxiety and reported confidence every four trials. **b. Rating variability.** Snapshot ratings varied between the two rating phases, with greater variability at intermediate than extreme anxiety ratings. **c. Choices.** The left snapshot was chosen more frequently when it had a higher relative average anxiety rating. **d. Choice consistency and rating difference.** Choice consistency increased with the absolute difference between the mean ratings of the two options. **e. Choice consistency and rating variability.** Choice consistency decreased with the summed variability of the two options. **f. Predicting choice consistency.** Logistic regression coefficients predicting choice consistency from summed ratings, rating difference and summed variability. **g. Model comparison.** Adding summed variability to a model already including rating difference significantly improved prediction of choice consistency. **h. Participant-level variability and choice sensitivity.** Mixed-effects logistic regression slopes for rating difference are plotted against rating variability per participant. Participants with higher overall rating variability showed weaker sensitivity to rating differences. Boxplot centre lines show medians; boxes show 25th-75th percentiles; whiskers extend 1.5× IQR; outliers are shown as points. Blue indicates empirical data and red indicates analyses repeated on data generated from the efficient coding model. Asterisks indicate statistical significance (*** $p < 0.001$).

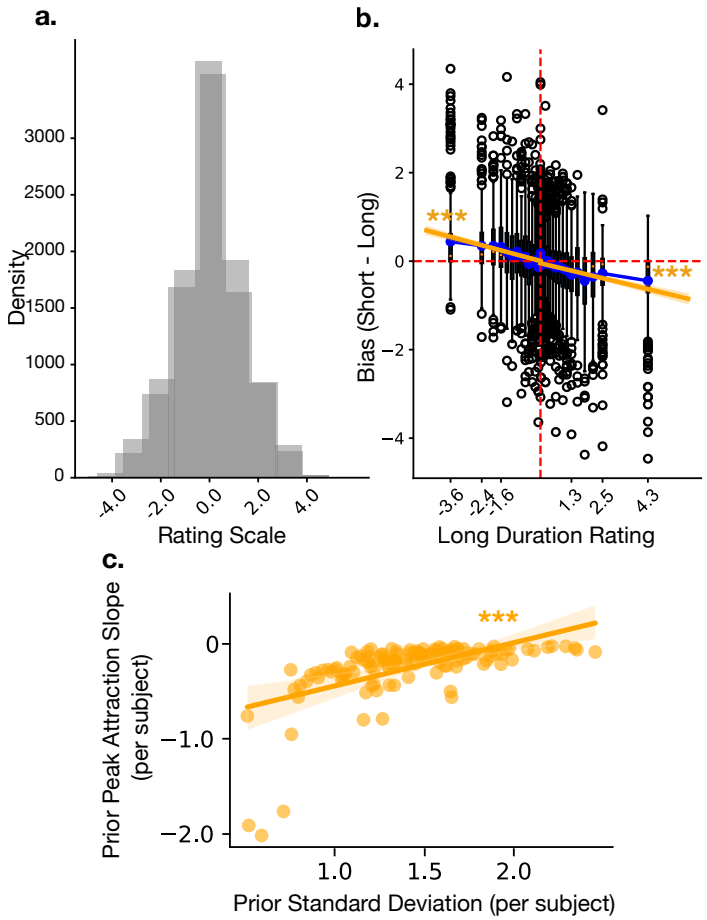

Fig. S11 Replication of Bayesian inference effects in Study 4 (N=120). **a. Prior distribution.** Aggregated anxiety rating distribution, z-scored per participant so that the participant-specific mode, used as the prior peak, is aligned at $x = 0$. **b. Bayesian attraction.** Bias in anxiety judgements, quantified as the short-long rating difference, is plotted against the participant's z-scored long-duration rating. The vertical dashed line marks the prior peak and the horizontal dashed line marks no difference between short- and long-duration ratings. Blue points show binned means; black boxplots show the distribution of short-long differences within rating bins. Orange lines show piecewise linear fits ± 95% CI, demonstrating that short-duration ratings were pulled toward the prior peak when stimuli were far from that peak. **c. Prior width and Bayesian attraction.** Participant-level Bayesian attraction slopes, estimated from regressions of bias on long-duration ratings, are plotted against participants' prior standard deviation. Each point represents one participant. Participants with wider priors showed less negative attraction slopes, indicating weaker pull toward the prior peak. Boxplot centre lines show medians; boxes show 25th-75th percentiles; whiskers extend 1.5× IQR; outliers are shown as points. Asterisks indicate statistical significance (*** $p < 0.001$).

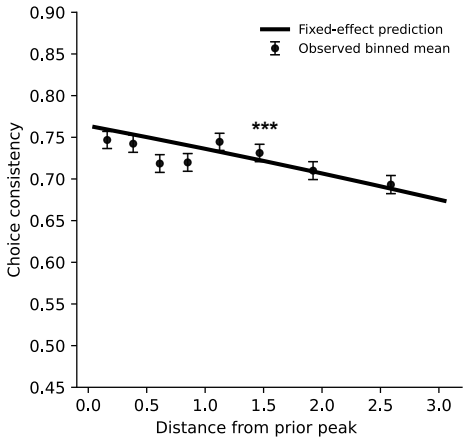

Fig. S12 Choice consistency as a function of continuous distance from the prior peak in Study 1 (N=57). Choice consistency is plotted against the distance of the choice-pair midpoint from the participant-specific prior peak. The black line shows the fixed-effect prediction from a logistic mixed-effects model controlling for the absolute rating difference between the two options, with random intercepts and random slopes for prior distance by participant. Points show observed binned means ± s.e.m. Choice consistency decreased as distance from the prior peak increased, consistent with higher discriminability near the prior peak. Asterisks indicate statistical significance (*** $p < 0.001$).

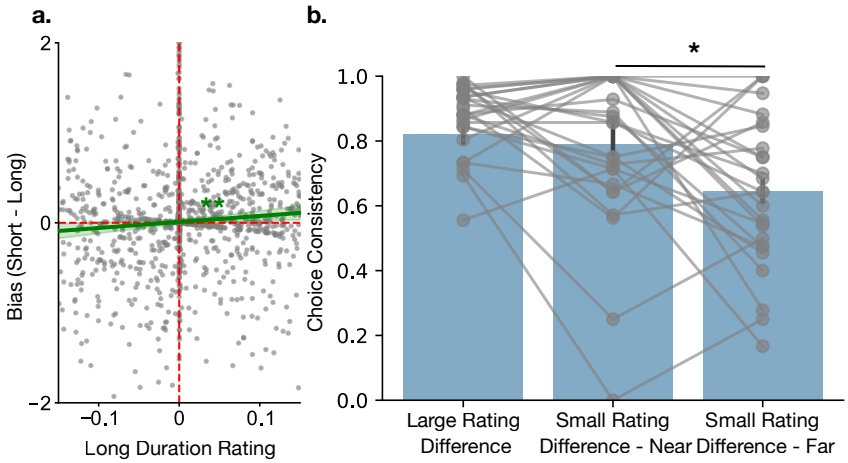

Fig. S13 Replication of efficient-coding effects in Study 4 (N=120). **a. Repulsive bias near the prior peak.** Bias in anxiety judgements, quantified as the short-long rating difference, is plotted against z-scored long-duration ratings close to the participant-specific prior peak. The green line shows a linear fit ± 95% CI, highlighting a repulsive bias near the prior peak. Red dashed lines indicate zero bias and the prior peak. **b. Choice consistency near and far from the prior peak.** Choice consistency is shown for trials with large rating differences, trials with small rating differences near the prior peak, and trials with small rating differences far from the prior peak. Grey lines link within-subject means; bars show group means ± 95% CI. Participants were more consistent for small-difference choices made near the prior peak than far from it. Asterisks indicate statistical significance (** $p < 0.01$).

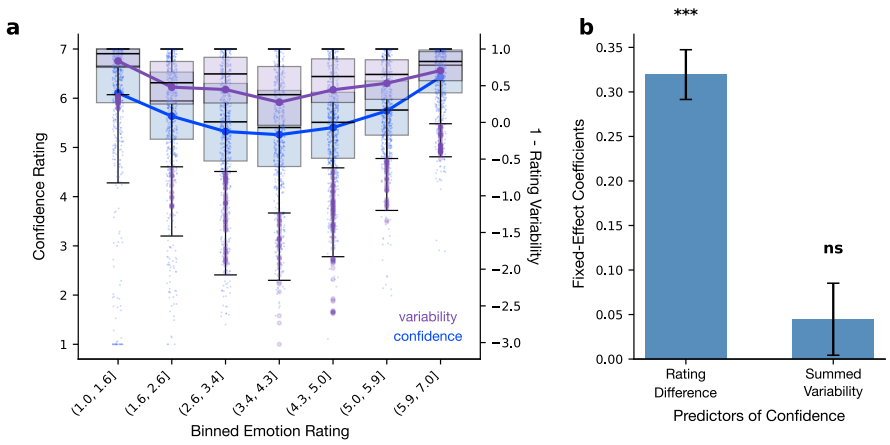

Fig. S14 Replication of awareness of uncertainty in Study 4 (N=120). **a. Confidence in emotion intensity ratings and rating variability.** Confidence ratings (blue; left axis) and rating reliability, quantified as 1−rating variability (purple; right axis), are plotted against binned anxiety ratings. For confidence, boxplots show the distribution across bins, overlaid with individual observations, and solid lines indicate the mean in each bin. For rating reliability, boxplots show the distribution across bins, outliers are shown individually in light purple, and solid lines indicate the mean in each bin. In all boxplots, the lower and upper hinges correspond to the 25th and 75th percentiles, and whiskers extend to 1.5× IQR. **b. Predicting confidence in choices.** Fixed-effect coefficients from a linear mixed-effects regression predicting confidence in emotion discrimination choices from rating difference and summed rating variability. A random intercept was included for each participant. Error bars indicate the standard error of the estimate. Asterisks indicate statistical significance (*** $p < 0.001$); n.s., not significant.

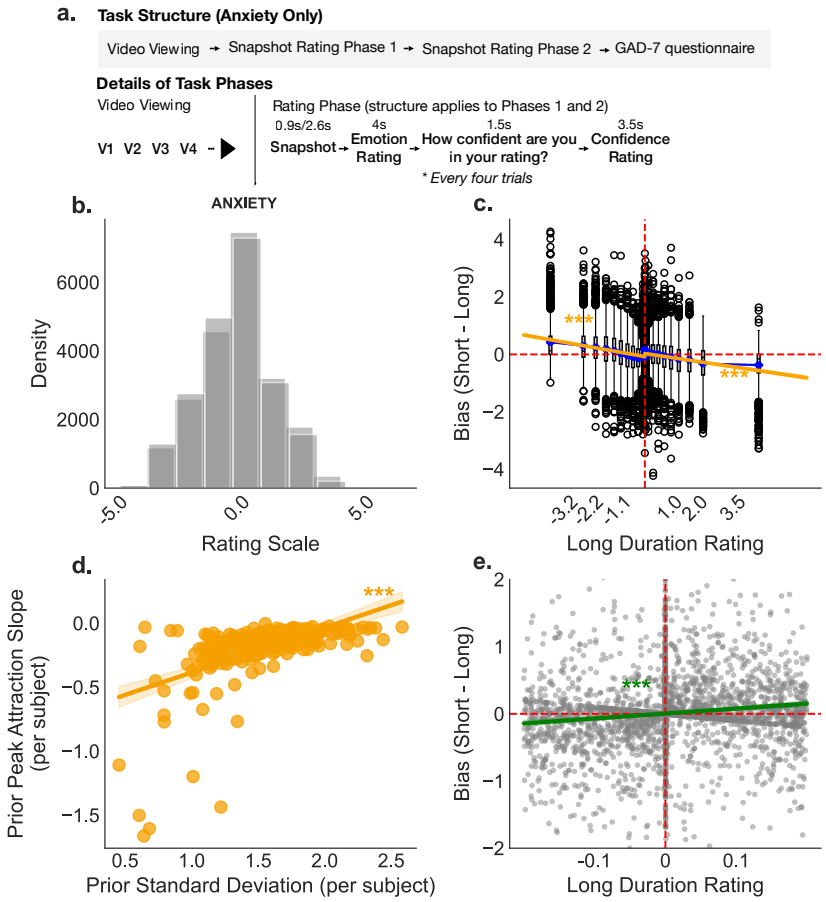

Fig. S15 Bayesian inference and efficient-coding effects in Study 5 (N=229). **a. Task structure.** Participants reporting anxiety symptoms completed an anxiety-only version of the task. They first viewed anxiety videos, then completed two snapshot rating phases, and finally completed the GAD-7 questionnaire. In each rating phase, participants rated anxiety elicited by previously viewed video snapshots and reported confidence every four trials. **b. Prior distribution.** Aggregated anxiety rating distribution, z-scored per participant so that the participant-specific mode, used as the prior peak, is aligned at $x = 0$. **c. Bayesian attraction.** Bias in anxiety judgements, quantified as the short-long rating difference, is plotted against the participant's z-scored long-duration rating. The vertical dashed line marks the prior peak and the horizontal dashed line marks no difference between short- and long-duration ratings. Blue points show binned means; black boxplots show the distribution of short-long differences within rating bins. Orange lines show piecewise linear fits ± 95% CI, showing attraction toward the prior peak for stimuli far from that peak. **d. Prior width and Bayesian attraction.** Participant-level Bayesian attraction slopes are plotted against participants' prior standard deviation. Each point represents one participant. Participants with wider priors showed less negative attraction slopes, indicating weaker pull toward the prior peak. **e. Repulsive bias near the prior peak.** Bias in anxiety judgements is plotted against z-scored long-duration ratings close to the participant-specific prior peak. The green line shows a linear fit ± 95% CI, highlighting a repulsive bias near the prior peak. Red dashed lines indicate zero bias and the prior peak. Asterisks indicate statistical significance (*** $p < 0.001$).

10 Supplementary information

References

[1] Barrett, L. F., Mesquita, B., Ochsner, K. N. & Gross, J. J. The Experience of Emotion. *Annual review of psychology* **58**, 373–403 (2007). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1934613/.

[2] Moors, A., Ellsworth, P. C., Scherer, K. R. & Frijda, N. H. Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review* **5**, 119–124 (2013). URL https://doi.org/10.1177/1754073912468165. Publisher: SAGE Publications.

[3] Wei, X.-X. & Stocker, A. A. A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience* **18**, 1509–1517 (2015). URL https://www.nature.com/articles/nn.4105. Publisher: Nature Publishing Group.

[4] Polanía, R., Woodford, M. & Ruff, C. C. Efficient coding of subjective value. *Nature Neuroscience* **22**, 134–142 (2019). URL https://www.nature.com/articles/s41593-018-0292-0. Number: 1 Publisher: Nature Publishing Group.

[5] Lazarus, R. S. Thoughts on the relations between emotion and cognition. *American Psychologist* **37**, 1019–1024 (1982). Place: US Publisher: American Psychological Association.

[6] Ekman, P. & Davidson, R. J. (eds) *The nature of emotion: Fundamental questions* The nature of emotion: Fundamental questions (Oxford University Press, New York, NY, US, 1994). Pages: xiv, 496.

[7] Frijda, N. H., Kuipers, P. & ter Schure, E. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology* **57**, 212–228 (1989). Place: US Publisher: American Psychological Association.

[8] Scherer, K. R. What are emotions? And how can they be measured? *Social Science Information* **44**, 695–729 (2005). URL https://doi.org/10.1177/0539018405058216. Publisher: SAGE Publications Ltd.

[9] Barrett, L. F. Solving the emotion paradox: categorization and the experience of emotion. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc* **10**, 20–46 (2006).

[10] Clore, G. L. & Ortony, A. in *Appraisal theories: How cognition shapes affect into emotion* 628–642 (The Guilford Press, New York, NY, US, 2008).

[11] Etkin, A., Büchel, C. & Gross, J. J. The neural bases of emotion regulation. *Nature Reviews. Neuroscience* **16**, 693–700 (2015).

[12] Wager, T. D. *et al.* A Bayesian model of category-specific emotional brain responses. *PLoS computational biology* **11**, e1004066 (2015).

[13] Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences* **114**, E7900–E7909 (2017). URL https://www.pnas.org/doi/10.1073/pnas.1702247114. Publisher: Proceedings of the National Academy of Sciences.

[14] Huys, Q. J. M. & Renz, D. A Formal Valuation Framework for Emotions and Their Control. *Biological Psychiatry* **82**, 413–420 (2017).

[15] Emanuel, A. & Eldar, E. Emotions as computations. *Neuroscience & Biobehavioral Reviews* **144**, 104977 (2023). URL https://www.sciencedirect.com/science/article/pii/S0149763422004663.

[16] Weidman, A. C., Steckler, C. M. & Tracy, J. L. The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion (Washington, D.C.)* **17**, 267–295 (2017).

[17] Ekman, P. Are there basic emotions? *Psychological Review* **99**, 550–553 (1992).

[18] Ortony, A. Are All "Basic Emotions" Emotions? A Problem for the (Basic) Emotions Construct. *Perspectives on Psychological Science* **17**, 41–61 (2022). URL https://doi.org/10.1177/1745691620985415. Publisher: SAGE Publications Inc.

[19] Scherer, K. R. & Moors, A. The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology* **70**, 719–745 (2019). Place: US Publisher: Annual Reviews.

[20] Attneave, F. Some informational aspects of visual perception. *Psychological Review* **61**, 183–193 (1954). Place: US Publisher: American Psychological Association.

[21] Barlow, H. B. in *Possible Principles Underlying the Transformations of Sensory Messages* (ed.Rosenblith, W. A.) *Sensory Communication* 0 (The MIT Press, 1961). URL https://doi.org/10.7551/mitpress/9780262518420.003.0013.

[22] Ganguli, D. & Simoncelli, E. P. Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation* **26**, 2103–2134 (2014).

[23] Laughlin, S. B. & Hardie, R. C. Common strategies for light adaptation in the peripheral visual systems of fly and dragonfly. *Journal of comparative physiology* **128**, 319–340 (1978). URL https://doi.org/10.1007/BF00657606.

[24] Knill, D. C. & Richards, W. (eds) *Perception as Bayesian Inference* (Cambridge University Press, Cambridge, 1996). URL https://www.cambridge.org/core/books/perception-as-bayesian-inference/0442F577F5E4CD874FA6819978574C8F.

[25] Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annual Review of Neuroscience* **24**, 1193–1216 (2001).

[26] Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002). URL https://www.nature.com/articles/415429a. Publisher: Nature Publishing Group.

[27] Doya, K. *Bayesian brain: Probabilistic approaches to neural coding* (MIT press, 2007).

[28] Palmer, J., Huk, A. C. & Shadlen, M. N. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision* **5**, 1 (2005). URL https://doi.org/10.1167/5.5.1.

[29] Ratcliff, R. & McKoon, G. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural computation* **20**, 873–922 (2008). URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2474742/.

[30] Huk, A. C., Palmer, J. & Shadlen, M. N. Temporal integration of visual motion information: Evidence from response times. *Journal of Vision* **2**, 228–228 (2002). URL https://jov.arvojournals.org/article.aspx?articleid=2120308. Publisher: The Association for Research in Vision and Ophthalmology.

[31] Stankevicius, A., Huys, Q. J. M., Kalra, A. & Seriès, P. Optimism as a Prior Belief about the Probability of Future Reward. *PLOS Computational Biology* **10**, e1003605 (2014). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003605. Publisher: Public Library of Science.

[32] Stuke, H., Weilnhammer, V. A., Sterzer, P. & Schmack, K. Delusion proneness is linked to a reduced usage of prior beliefs in perceptual decisions. *Schizophrenia Bulletin* **45**, 80–86 (2019). Place: United Kingdom Publisher: Oxford University Press.

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

[34] Dejonckheere, E. *et al.* Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment* **34**, 1138–1154 (2022). Place: US Publisher: American

Psychological Association.

[35] Hoemann, K. *et al.* Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific Reports* **10**, 12459 (2020). URL https://www.nature.com/articles/s41598-020-69180-y. Publisher: Nature Publishing Group.

[36] Ong, D. C., Zaki, J. & Goodman, N. D. Affective cognition: Exploring lay theories of emotion. *Cognition* **143**, 141–162 (2015).

[37] Anzellotti, S., Houlihan, S. D., Liburd, S. & Saxe, R. Leveraging facial expressions and contextual information to investigate opaque representations of emotions. *Emotion (Washington, D.C.)* **21**, 96–107 (2021).

[38] Goel, S., Jara-Ettinger, J., Ong, D. C. & Gendron, M. Face and context integration in emotion inference is limited and variable across categories and individuals. *Nature Communications* **15**, 2443 (2024). URL https://www.nature.com/articles/s41467-024-46670-5. Publisher: Nature Publishing Group.

[39] Hoemann, K., Gendron, M. & Barrett, L. F. Mixed emotions in the predictive brain. *Current opinion in behavioral sciences* **15**, 51–57 (2017). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5669377/.

[40] Barrett, L. F., Mesquita, B. & Gendron, M. Context in Emotion Perception. *Current Directions in Psychological Science* **20**, 286–290 (2011). URL https://doi.org/10.1177/0963721411422522. Publisher: SAGE Publications Inc.

[41] Kafetsios, K. & Hess, U. Personality and the accurate perception of facial emotion expressions: What is accuracy and how does it matter? *Emotion (Washington, D.C.)* **22**, 100–114 (2022).

[42] Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 1310–1321 (2012). URL https://doi.org/10.1098/rstb.2011.0416.

[43] Boundy-Singer, Z. M., Ziemba, C. M. & Goris, R. L. T. Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour* **7**, 142–154 (2023). URL https://www.nature.com/articles/s41562-022-01464-x. Publisher: Nature Publishing Group.

[44] Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience* **19**, 366–374 (2016). URL https://www.nature.com/articles/nn.4240. Publisher: Nature Publishing Group.

[45] Lee, H.-H., Liu, G. K.-M., Chen, Y.-C. & Yeh, S.-L. Exploring quantitative measures in metacognition of emotion. *Scientific Reports* **14**, 1990 (2024). URL https://www.nature.com/articles/s41598-023-49709-7. Publisher: Nature Publishing Group.

[46] Plate, C. R. *et al.* Computational characterization of metacognitive ability in subjective decision-making. *bioRxiv: The Preprint Server for Biology* 2025.05.23.655775 (2025).

[47] Kiani, R., Corthell, L. & Shadlen, M. N. Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron* **84**, 1329–1342 (2014). URL https://www.sciencedirect.com/science/article/pii/S0896627314010964.

[48] Dunn, B. D., Dalgleish, T., Lawrence, A. D., Cusack, R. & Ogilvie, A. D. Categorical and Dimensional Reports of Experienced Affect to Emotion-Inducing Pictures in Depression. *Journal of Abnormal Psychology* **113**, 654–660 (2004). Place: US Publisher: American Psychological Association.

[49] Punkanen, M., Eerola, T. & Erkkilä, J. Biased emotional recognition in depression: Perception of emotions in music by depressed patients. *Journal of Affective Disorders* **130**, 118–126 (2011). URL https://www.sciencedirect.com/science/article/pii/S0165032710006488.

[50] Hoven, M. *et al.* Abnormalities of confidence in psychiatry: an overview and future perspectives. *Translational Psychiatry* **9**, 268 (2019). URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6803712/.

[51] Seow, T. X. F., Fleming, S. M. & Hauser, T. U. Metacognitive biases in anxiety-depression and compulsivity extend across perception and memory. *PLOS Mental Health* **2**, e0000259 (2025). URL https://journals.plos.org/mentalhealth/article?id=10.1371/journal.pmen.0000259. Publisher: Public Library of Science.