Common psychiatric treatments alter affective dynamics

Quentin Dercon[1] ✉, Quentin J. M. Huys[1], Robb B. Rutledge[2,3,4], and Camilla L. Nord[5,6]

[1]Applied Computational Psychiatry Lab, Mental Health Neuroscience Department, Division of Psychiatry and Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Department of Imaging Neuroscience, Queen Square Institute of Neurology, UCL, UK
[2]Department of Psychology, Yale University, New Haven, CT, USA
[3]Wu Tsai Institute, Yale University, New Haven, CT, USA
[4]Department of Psychiatry, Yale University, New Haven, CT, USA
[5]MRC Cognition and Brain Sciences Unit, University of Cambridge, UK
[6]Department of Psychiatry, University of Cambridge, UK

✉Address correspondence to: quentin.dercon.22@ucl.ac.uk.

Abstract

Affective states are dynamic, fluctuating in response to recent events: an unexpected pleasure, a disappointing loss. Affective biases, which cause disruptions in these dynamics, are core components of mental ill-health, but the specific effects of treatments on these biases are poorly understood. Here, we investigate the impact of common psychiatric treatments on subjective assessments of happiness, confidence, and engagement during a reinforcement learning task (N=935; 130 taking antidepressant medications). Half (N=459) of the participants were randomised to practice a common psychotherapeutic technique—cognitive distancing—throughout the task. From a joint computational model of learning and affect, we find evidence for distinct and overlapping impacts of psychiatric treatments on affective dynamics. Cognitive distancing attenuates downward drift in happiness and engagement and increases recency bias in the affective impact of recent choices. Conversely, antidepressant use increases baseline happiness and confidence in individuals with similar levels of current symptoms, and decreases recency bias such that more past events influence affective states. Crucially, both cognitive distancing and antidepressant use converge to dampen negative biases in happiness and confidence specifically in participants experiencing higher levels of anxiety and depression symptoms. Together, our results indicate that common treatments for mental ill-health may alter symptoms through their impact on affective dynamics, but via distinct mechanisms.

1 Background

How are you feeling right now? Research across economics, psychology, and health sciences suggests the answer to this question—your subjective well-being—is closely tied to objective quality of life[1,2] and health across the lifespan[3]. But feelings are far from static, momentarily fluctuating in response to recent events[4–6], and even individual choices. Frequently asking participants to rate their feelings enables a read-out of moment-to-moment changes in subjective well-being, or their *affective dynamics*.

In influential work, Rutledge *et al.* (2014)[5] demonstrated momentary happiness ratings during a gambling task could be accurately predicted by a computational model incorporating the average reward for a gamble (*expected value*) and the outcome of the gamble minus this average (*prediction error*). Using a task where reward magnitude and probability were uncorrelated, Blain & Rutledge (2020)[7] subsequently showed that momentary happiness is particularly sensitive to changes in learning-related variables—specifically, prediction errors for reward probability—as compared to as compared to reward information that was relevant to behaviour but not learning. Links between happiness ratings and learning-related quantities may extend to subjective assessments of other affective states. Theoretical accounts posit that momentary confidence is the approximate probability a choice is correct[8,9] (though see[10,11]), while effort costs decrease the value of choices independently of reward probability[12,13], in turn influencing momentary engagement. Together, these results suggest that affective dynamics are closely coupled with objective quantities that drive choices during learning.

Biases in subjective affect are a core feature of mental ill-health. Symptoms of depression have been consistently linked to lower[7,14] and less stable[15,16] momentary happiness, while transdiagnostic features of mental ill-health have been linked to biased confidence judgements at different timescales[17–20] and impairments in motivation and engagement[13,21–23]. Affective biases maintain symptoms of mental ill-health by inducing changes in emotion processing and perception[24,25]. For example, negatively biased perception—a common feature of depression[26]—may cause low mood by making outcomes appear less rewarding; low mood in turn further negatively biases perception, causing a positive feedback loop which

spirals toward a depressive episode[27]. Successful psychiatric treatment may act to perturb these maladaptive cycles. Short-term selective serotonin reuptake inhibitor (SSRI) administration induces positive perceptual biases in healthy participants[28], suggesting that affective biases may be an early target of antidepressant drugs, acting to shift choices away from those that maintain low mood[29]. Crucially, given that affective biases are precipitated and maintained by negative thinking patterns—a core target of cognitive psychotherapy[30,31]—they may represent a transdiagnostic treatment target of psychological and pharmacological interventions for symptoms of mental ill-health.

Here, we aimed to link choice behaviour to affective dynamics throughout a reinforcement learning (RL) task[32–34] (Figure 1A) and to relate this to mental ill-health symptoms and treatments. We asked online participants (N=935) to rate their feelings (from 0-100) on one of three different affect scales—happiness, confidence, and engagement—after receiving feedback on each choice they made. Half (49.1%) of the participants were randomised to an acute psychological intervention known as "cognitive distancing", a common[35] and effective[31] component of psychological therapy which alters learning in this task[34]. We also collected information on current antidepressant medication use in a demographic questionnaire (reported by 13.9% of participants), and derived transdiagnostic mental health symptom factor scores from a psychiatric questionnaire battery[36,37]. We then assessed how participants' affect ratings across each of the three scales covaried with learning-related outcomes throughout the task, accounting for underlying affective biases, using computational modelling. By quantifying the associations between model-derived measures of affective dynamics and transdiagnostic features of mental ill-health, the cognitive distancing intervention, and self-reported antidepressant medication use, we asked whether affective dynamics may be a common target of treatments for symptoms of mental ill-health.

2 Methods

2.1 Online experiment and sample

A total of 995 participants were recruited via Prolific[38] over three weeks in April-May 2021. Participants were recruited in batches with fixed pre-screeners for age range, gender, and history of any mental health diagnosis, which resulted in a sample broadly representative of the UK population in terms of these characteristics (see[34] for details). After completing a demographic questionnaire, which included questions on current medication usage and mental health diagnoses, participants completed the RL task described below. They then took a short test of working memory (visual digit span), and answered questions from several psychiatric questionnaires, answers from which were used to derive three validated transdiagnostic features of mental health (anxiety/depression, compulsive behaviour, and social withdrawal)[36], using methods for computational factor modelling[37] described in detail elsewhere[34,39]. Sixty participants were excluded for meeting pre-registered criteria[34]. The study was approved by the University of Cambridge Human Biology Research Ethics Committee (HBREC) (HBREC.2020.40) and jointly sponsored by the University of Cambridge and Cambridge University Hospitals National Health Service (NHS) Foundation Trust (IRAS ID 289980). All participants provided written informed consent through an online form, in line with University of Cambridge HBREC procedures for online studies.

2.2 Reinforcement learning task

The reinforcement learning task in the present study—the probabilistic selection task[32,33]—involved learning which symbol in each of three pairs was more likely correct. Consistent choice of the "better" symbol in each pair enabled a participant to accumulate more points (and maximise their chances of winning a monetary bonus). One symbol in each pair was always more likely correct, but the contingencies varied across the pairs, from 0.8/0.2 ('AB') to 0.7/0.3 ('CD'), to 0.6/0.4 ('EF'). All participants saw the same six symbols, but the pairs were randomised across individuals and counterbalanced across trials. After making a choice, participants received feedback ("Correct!" or "Incorrect."), and then rated their subjective happiness, confidence, or engagement (Figure 1A).

One of the three questions was asked after each trial outcome, with each question asked twenty times per block of sixty trials, and never more than twice in a row. Participants were also asked to rate (again from 0-100) how fatigued they felt compared to the beginning of the block after the end of each of the sixty-trial training blocks. After six training blocks, participants were tested on all fifteen unique pairs without feedback. We previously reported the effects of cognitive distancing on task performance and learning, including results of the test phase, in the same sample[34].

2.3 Acute psychological intervention and antidepressant use

Half of the participants (n=459; 49.1%) were randomised to be taught, and then practice throughout the task, a psychotherapeutic technique termed "cognitive distancing", which encouraged them to "take a step back" from their emotional reactions to feedback throughout the task (see here[34] for further details). Apart from an additional instructional video before the task started and a small prompt to *"Distance yourself..."* which appeared with each fixation cross (Figure 1A), the task was identical for distanced and non-distanced participants. To explore similarities between the effects of this psychological intervention, and a pharmacological treatment for mental ill-health, antidepressant medication, we also asked participants to report their current medication use: 130 participants (13.9%) reported current antidepressant use, with the majority (n=94; 72.3%) taking an SSRI.

2.4 Computational modelling: joint RL-affect models

For consistency with previous literature, we used the well-characterised model of momentary happiness first described by Rutledge *et al.* (2014)[5] as a baseline model. This model assumes fluctuations around a baseline (i.e., longer-term mean) can be captured by a weighted sum of recent expected values and prediction error and, importantly, does not condition on previous ratings.

The Rutledge *et al.* (2014)[5] model has been primarily validated in (e.g., gambling) tasks where expected values and prediction errors are explicitly available to participants[5,40]. As such, we extended it to account for learning in this task. Specifically, our model—which we term a joint RL-affect model—comprised two components: (1) a Q-learning model to infer

expected values and prediction errors from participants' choices (which has been shown to accurately capture choice behaviour in this task[33,34]); and (2) a model for momentary affect which assumes fluctuations around a baseline can be captured by a recency-weighted sum of Q-learning model-derived expected values and prediction errors[5,41]. Hierarchical models were simultaneously fitted to task choices plus happiness, confidence, and engagement self-reports, assuming different parameter weights across all scales and participants.

2.4.1 RL models

A Q-learning model infers expected values (termed Q-values) from participant choices—the action (a_t) of choosing one symbol over the other in each of the three pairs (denoted as states, s_t)—by assuming they update at each trial t based on prediction errors δ_t, with the update magnitude controlled by a learning rate $\alpha \in [0, 1]$:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha\delta_t \quad \text{where} \quad \delta_t = r_t - Q_t(s_t, a_t). \tag{1}$$

We additionally considered a dual learning rate Q-learning model for choices, in which the learning rate is assumed to differ depending on whether the outcome was rewarding (α_{reward}) or not (α_{loss})[33]:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_{\text{reward}}[\delta_t]_+ + \alpha_{\text{loss}}[\delta_t]_-. \tag{2}$$

In both cases, the difference in Q-values between the chosen (a_t) and avoided action ($\bar{a}_t$) is transformed to a choice probability using a softmax function, weighted by an inverse temperature β:

$$P_t(s_t, a_t) = \left(1 + \exp\left\{-\beta\left[Q_t(s_t, a_t) - Q_t(s_t, \bar{a}_t)\right]\right\}\right)^{-1}. \tag{3}$$

2.4.2 Affect models

Baseline model

Affect ratings scaled to $[0, 1]$ for each participant and rating type (i.e., happiness, confidence or engagement) were assumed to be drawn from independent Beta distributions with a

mean-variance reparameterization which models the shape parameters as functions of a (conditional) mean and precision[42]. Extreme ratings (0 or 1) were allowed in the task, so we transformed ratings to the $(0, 1)$ interval using a simple transformation[43] ($y' = \frac{y(N-1)+0.5}{N}$; here N is the total number of participants but could be any large number). A logit link function was used, so the Beta regression weights (excluding the intercept w_0^p) can be interpreted as the log odds ratio per unit change in the covariate for an increase in affect rating, with other terms held constant[a]. Following[5,41] we write the ith participant's affect rating $Y^{(i)}$ for rating type p at trial t as,

$$Y_{t,p}^{(i)} \sim \text{Beta}\Big(\mu_t^{(i)}\phi_p, (1-\mu_t^{(i)})\phi_p\Big) \tag{4}$$

$$\log\left(\frac{\mu_t^{(i)}}{1-\mu_t^{(i)}}\right) = w_0^p + w_2^p \sum_{t'=1}^{t} \gamma_p^{t-t'}\, Q_{t'}(s_{t'}, a_{t'}) + w_3^p \sum_{t'=1}^{t} \gamma_p^{t-t'}\, \big[r_{t'} - Q_{t'}(s_{t'}, a_{t'})\big], \tag{5}$$

where t is the overall trial number at rating number i for rating type p, γ is the discount or 'forgetting' factor which imposes a strict weighting on recent trials, and w_k^p the weights on each of the k quantities of interest for the pth rating type; and $Q_{t'}$ and $r_{t'}$ are the Q-learning model-derived expected values for the chosen symbol $a_{t'}$ in the state $s_{t'}$ (i.e., the pair of symbols presented on trial t') and the feedback valence (± 1 for correct/incorrect to allow for negative Q-values) respectively, both at trial t'. Note that both sums are over all previous trials, not just those of rating type p.

Accounting for drift over time

We also fit models including an extra weight w_1^p to account for potential "drift over time" in affect[44], by modifying (5) as follows,

$$\log\left(\frac{\mu_t^{(i)}}{1-\mu_t^{(i)}}\right) = w_0^p + w_1^p\, \tau_t + w_2^p \sum_{t'=1}^{t} \gamma_p^{t-t'}\, Q_{t'}(s_{t'}, a_{t'}) + w_3^p \sum_{t'=1}^{t} \gamma_p^{t-t'}\, \big[r_{t'} - Q_{t'}(s_{t'}, a_{t'})\big], \tag{6}$$

where τ_t is some measure of time elapsed up to trial t: either trial number, block number, overall time elapsed, or time elapsed since the start of that block (see Model comparison).

[a] Note that our approach differs slightly from that of Rutledge *et al.* (2014)[5], who assume momentary happiness ratings follow a Gaussian distribution; see Forbes & Bennett (2024)[41] for a validation of this Beta regression approach.

2.4.3 Fits to data

Models were fitted in a hierarchical Bayesian manner, with approximate posteriors derived via automatic differentiation variational inference (ADVI)[45] implemented in CmdStan[46]. All models were fit to choices and ratings on all three affect scales simultaneously across both distancing and non-distancing participants, with separate weights and decay factors assumed for each person and question, and separate group-level (hyper)priors on each parameter. In other words, participant-level parameter distributions are assumed to be conditionally independent given the group-level distribution over that parameter. Individual-level predictive accuracy was assessed by comparing responses predicted from each participant's approximate posterior to their observed affect ratings via pseudo-R^2, defined following previous work[42] as the squared correlation between observed and mean posterior predictions.

2.4.4 Model comparison

In the affect model, we tested for "drift over time"[44]; and in the Q-learning model we tested for separate learning rates for rewarding and non-rewarding outcomes. We assumed the Rutledge *et al.* (2014)[5] model (equation 5) to be the baseline model, and so the parameters in this model were included in all models.

Drift over time in affect may be particularly relevant to our task, as participants were able to take as long as they wished to rate their subjective feelings, and time between blocks was additionally unconstrained. As such, we compared four models with different measures of time elapsed (i.e., variants of equation 6) to the baseline model (equation 5), and either single or dual learning rates. The extra parameter added linear weights on either trial number, block number, or total time elapsed. We also tested a final model with two extra parameters: weights on both total time elapsed and time elapsed since the beginning of that block. We then compared all ten models in terms of their approximate leave one out (LOO) expected log pointwise predictive density (ELPD), a metric of estimated out-of-sample predictive accuracy[47], corrected for the use of variational approximations to the true posterior[48].

2.5 Statistical analysis

For consistency with computational modelling, we adopt a fully Bayesian approach for statistical analyses where possible. As such, results are given as estimates with a highest density interval (HDI), which (unlike a confidence interval (CI)[49]) can be interpreted as the probability that the true value falls within a given range. We report 95% HDIs to align with convention but interpret results in terms of strength of evidence throughout; an overlap with the null value should not be seen as evidence for lack of an effect, but rather weakened evidence for it.

2.5.1 Associations between model parameters and mental health symptoms & treatments

We tested the effect of differences in transdiagnostic mental health symptoms, current self-reported antidepressant use, or cognitive distancing on model parameters using generalised linear models (GLMs), adjusted for age, gender, and working memory capacity (measured with visual digit span), separately for each rating type. GLMs relating model parameters to antidepressant use were run with and without adjustment for concurrent anxiety/depression symptoms (i.e., factor score), as medication use was not randomised. We also considered whether the effect of cognitive distancing and current antidepressant use on affect model parameters may differ in relation to their association with transdiagnostic mental health, by including factor score as an interaction term in GLMs.

Posterior samples for GLM coefficients were obtained via Markov chain Monte Carlo (MCMC) implemented in CmdStan[46], with 2,000 warm-up and 10,000 sampling iterations for each of four chains, using models and priors from the `rstanarm` R package[50]. Response distributions were assumed Gaussian for all parameters except for learning and decay rates, which were modelled via Gamma family GLMs (with log link functions). See Interpretation of model-derived parameters in the Supplementary Methods for intuition as to how these regression coefficients are interpreted.

2.5.2 Differences in associations with baseline affect between transdiagnostic factors

To account for potential collinearities in the three transdiagnostic mental health symptom factor scores obtained via computational factor modelling, we used partial least squares

(PLS) regression to test which of the transdiagnostic factor scores was most strongly associated with baseline affect (w_0^p). PLS regression is a data-driven method which identifies latent components linking multivariate responses to predictors based on shared covariance, and so is well-suited to the problem of multicollinearity[51].

In line with best-practices[39], we first identified the number of components that best described our data in a training set (80% of participants) in terms of mean squared error (MSE) and R^2 via ten-fold cross-validation. We then validated the predictive accuracy of this number of components in held-out test data (20% of participants), and formally tested this using a permutation test, where the PLS regression model was re-trained on 10,000 training datasets with shuffled outcome labels (providing a null distribution), and the fraction of these datasets where the MSE between the test data and the predictions from the permuted datasets was lower than the true train-test MSE taken as the p-value. The PLS regression with the chosen number of components was then refitted in the whole dataset, to obtain component loadings on each of the responses and predictors. Lastly, we computed bias-corrected and accelerated (BCa) CIs for each of the loadings, plus the differences in loadings between transdiagnostic factors, from 10,000 bootstrap samples.

3 Results

3.1 A computational model of subjective happiness accounting for learning and affective drift also captures momentary confidence and engagement

Following model comparison, we found that the best-fitting RL-affect model included separate learning rates for rewarding and non-rewarding outcomes and a linear effect of time elapsed since the beginning of the task (equation 6; Figure 1D). This model, when fitted to all affect ratings and participants simultaneously, explained participants' variance in happiness, confidence, and engagement assessments with similar accuracy (mean [standard deviation (SD)] pseudo-R^2 = 0.40-0.42 [0.23-0.26]; see Figure 1C for example individuals).

3.2 Baseline affect is negatively associated with transdiagnostic mental health

We first assessed whether individuals' estimated model parameters from the best-fitting joint RL-affect model were associated with transdiagnostic features of mental health.

In line with previous work[7,14], we found strong evidence for a negative association between baseline happiness (w_0^{happy}) and anxiety/depression symptoms from a linear model adjusting for age, gender, digit span, and distancing (mean 4.44-point lower baseline happiness rating per SD increase in factor score; 95% HDI = [-5.46, -3.41]). Higher anxiety/depression factor scores were additionally associated with lower baseline confidence (mean 3.75-point lower $w_0^{\text{confident}}$ per SD increase in score; 95% HDIs = [-5.01, -2.46]) and lower baseline engagement (mean 2.90-point lower w_0^{engaged} per SD increase in score; 95% HDIs = [-4.12, -1.66]; Figure 2A*i*). We also found that higher anxiety/depression factor scores were associated with lower odds of increases over time in happiness (mean 11.2% lower w_1^{happy} per SD increase in anxiety/depression score per hour; 95% HDI for multiplier = [0.808, 0.971]) and engagement (17.0% lower w_1^{engaged} for a unit increase in anxiety/depression score per hour; 95% HDI for multiplier = [0.726, 0.943]; Figure 2A*ii*), suggesting a higher rate of decline (i.e., more drift) in happiness and engagement in participants with higher anxiety/depression scores.

Higher compulsive behaviour and social withdrawal factor scores were also each associated with lower baseline happiness (e.g., mean 2.20-point lower w_0^{happy} [95% HDI = (-3.31,

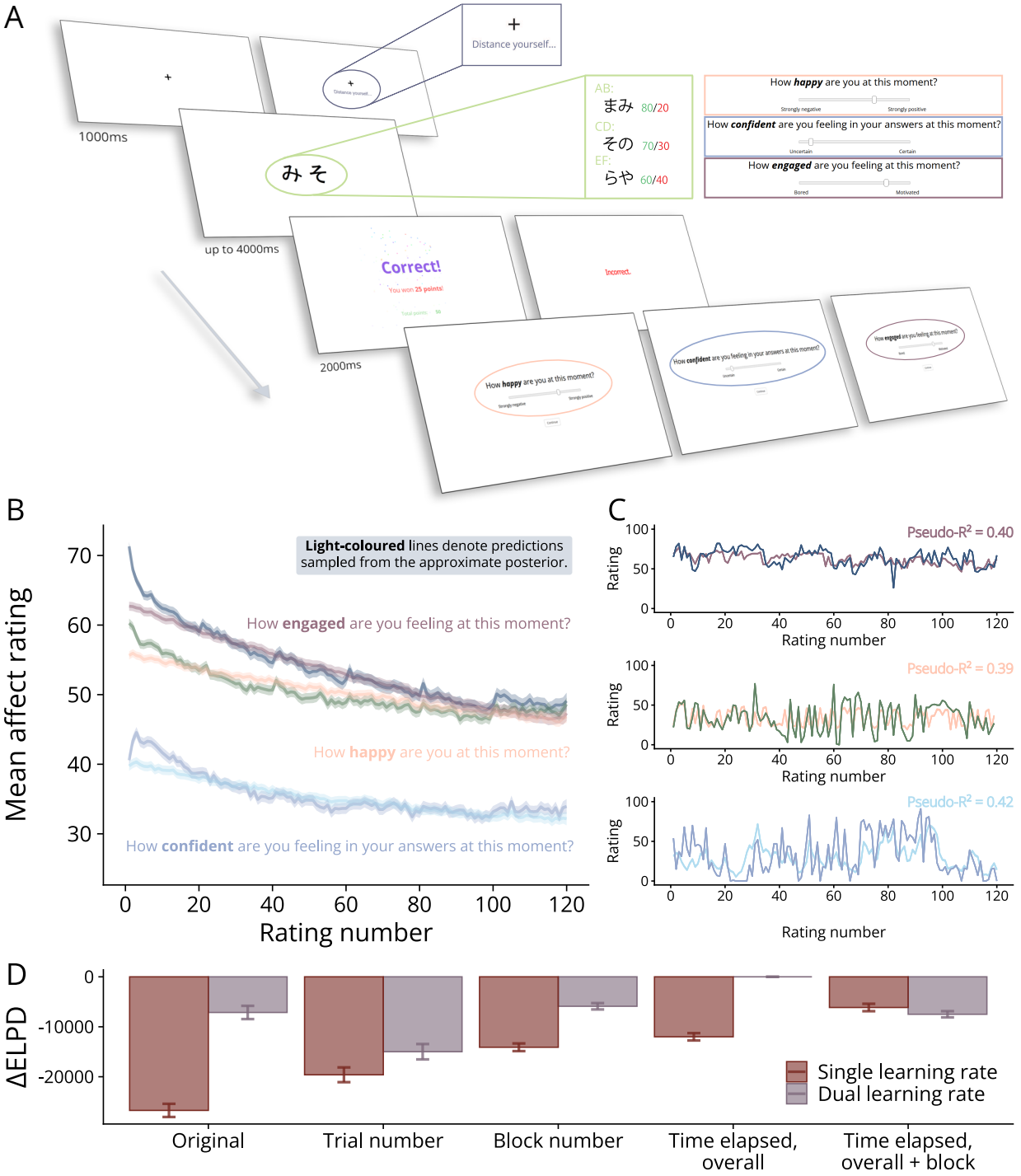

Figure 1: **Task design, affect model posterior predictions and model comparison.**. **A**. The task design. **B**. Mean affect ratings for each rating type (engaged, happy, or confident), by distancing group, compared to model predictions (light-coloured lines). **C**. Example comparison of predictions from the best-fitting model (light-coloured lines) to raw affect ratings from three different individuals with the median pseudo-R^2 for each rating type. **D**. Model fit compared to the best-fitting model (time elapsed, overall with dual learning rate) in terms of their ELPD (i.e., higher ELPD [or less negative ELPD compared to the best-fitting model] is better), estimated via Bayesian approximate LOO cross-validation[47]. Ribbons in B-C and error bars in D denote standard errors.

-1.06)] per SD increase in compulsive behaviour score; Figure Figure 2B*i*) and confidence (e.g., mean 3.59-point lower $w_0^{\text{confident}}$ [95% HDI = (-3.98, -1.38)] per SD increase in social withdrawal score; Figure 2C*i*), but were not associated with greater drift in affect over time (Figure 2B*ii* & Figure 2C*ii*). Associations between transdiagnostic symptom scores and choice-related affect measures (i.e., w_2^p and w_3^p were also observed (Figure S3), as were strong associations between post-block fatigue ratings and both baseline and drift in affect over time (Figure S5; see Supplementary Results for more details).

3.3 Baseline affect is most strongly associated with anxiety/depression symptoms

The three transdiagnostic mental health symptom factor scores obtained via computational factor modelling were highly correlated (r = 0.47 [95% CI: (0.42, 0.52)] between anxiety/ depression and compulsive behaviour; r = 0.61 [95% CI: (0.57, 0.65)] between anxiety/ depression and social withdrawal; r = 0.42 [95% CI: (0.37, 0.47)] between compulsive behaviour and social withdrawal). As such, to compare the strength of associations between baseline affect and transdiagnostic symptom factors, we used a partial least squares regression model to relate baseline affect to the three scores plus age, gender, digit span, and distancing.

We found that a three-component model represented the best compromise between predictive accuracy (in training data) and parsimony (Figure 2D), and there was strong statistical evidence that this model could accurately predict responses in held-out test data (permutation test $p<0.0001$). The first component of the model negatively loaded on baseline happiness (loading = -0.185, BCa bootstrapped 95% CI = [-0.217, -0.135]), confidence (loading = -0.112, BCa bootstrapped 95% CI = [-0.157, -0.049]), and engagement (loading = -0.131, BCa bootstrapped 95% CI = [-0.171, -0.068]) (Figure 2F). It also positively loaded on each of the transdiagnostic symptom factors: anxiety/depression (loading = 0.660, BCa bootstrapped 95% CI = [0.608, 0.733]), compulsive behaviour (loading = 0.491, BCa bootstrapped 95% CI = [0.415, 0.545]), and social withdrawal (loading = 0.577, BCa bootstrapped 95% CI = [0.529, 0.623]). The other two components did not show this pattern (Figure 2E). There was strong evidence that the first component's loading was higher for anxiety/depression than both compulsive behaviour and social withdrawal (component 1 loading difference [BCa bootstrapped 95% CI] = 0.169 [0.083, 0.299] and 0.083 [0.037, 0.152] respectively; Figure 2G).

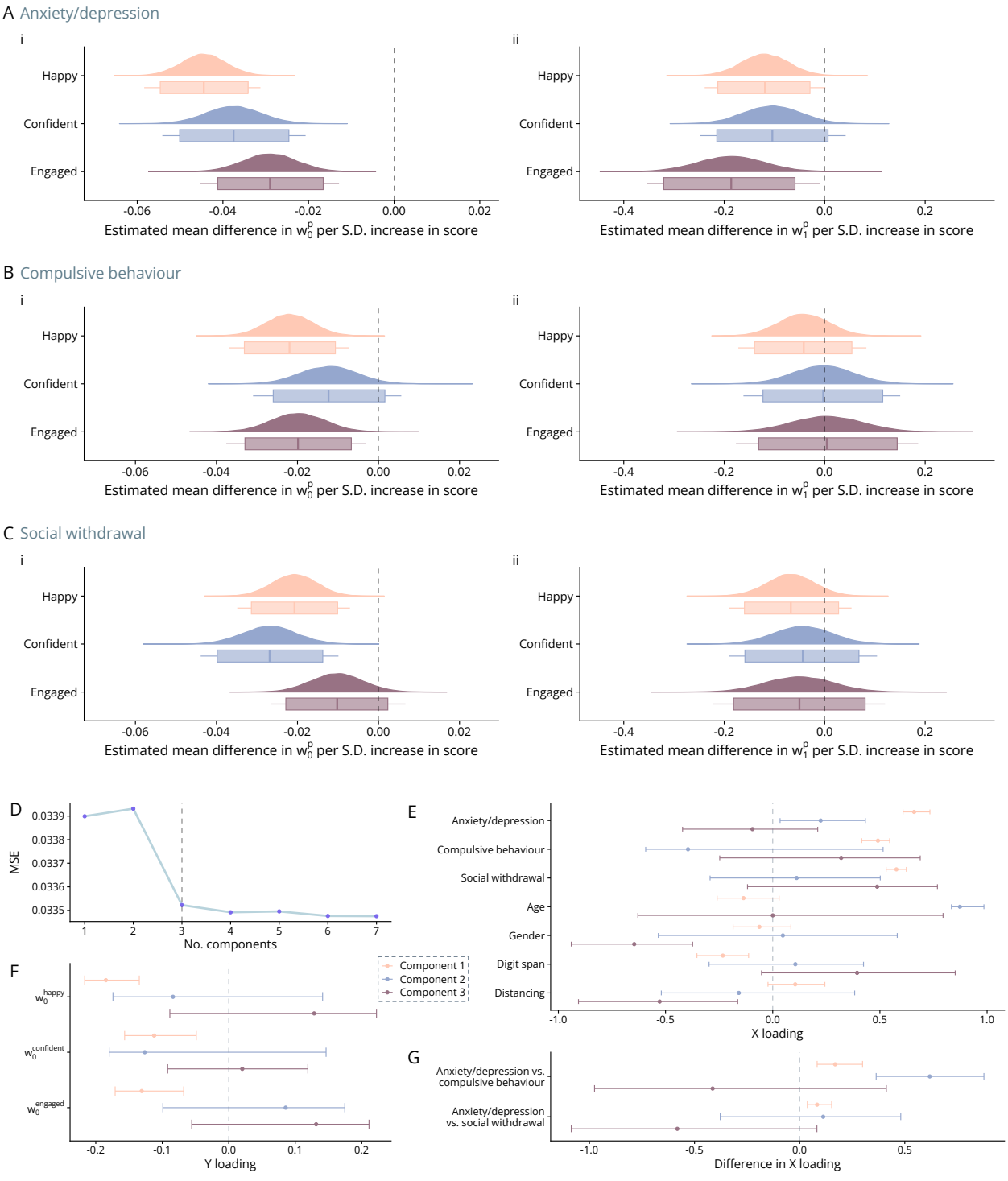

Figure 2: **Associations between affect parameters and transdiagnostic mental health dimensions, and results of PLS regression. A-C.** Estimated differences in baseline affect (w_0^p; *i*) or drift in affect over time (w_1^p; *ii*) for a unit increase in anxiety/depression (A), compulsive behaviour (B), or social withdrawal (C) transdiagnostic symptom factor score. **D-G.** Results of partial least squares regression: elbow plot of ten-fold cross-validated mean squared error for models with increasing numbers of components (D); loadings of the three-factor model on independent (E) and response (F) variables; and loading differences between anxiety/depression and other transdiagnostic factors (G). Boxplot boxes in A-C denote 95% HDIs and lines denote 99% HDIs; error bars in E-G denote BCa bootstrapped 95% CIs.

3.4 Cognitive distancing slows affective drift and antidepressant use positively modulates baseline affect

We then assessed the evidence for effects of cognitive distancing and antidepressant use on choice-independent affective dynamics: baseline affect and its drift.

Previously, in the same sample, we reported evidence from linear mixed models that participants practising cognitive distancing declined slightly less in happiness and engagement, but not confidence, over the course of the task[34]. Evidence from our RL-affect model was consistent with a decrease in affect drift over time: distancing individuals on average drifted less across the task in happiness (estimated mean 21.3% higher odds of increase in happiness over per hour; 95% HDI for multiplier = [1.01, 1.46]; Figure 3A*ii*), in spite of lower baseline engagement (estimated mean = -2.89 points; 95% HDI = [-5.42, -0.464]; Figure 3A*i*). There was also some weak evidence of less drift in engagement in participants randomised to the intervention (17.4% higher w_1^{engaged} in distancing individuals; 95% HDI for multiplier = [0.91, 1.52]; Figure 3A*ii*).

There was limited evidence for any difference in affective drift in participants taking antidepressants (Figure 3B*ii*). However, there was evidence that participants self-reporting current antidepressant use had 3.50-point higher baseline happiness (95% HDI = [0.311, 6.60]) and 2.69-point higher baseline confidence (with much weaker evidence: 95% HDI = [-1.16, 6.54]), after adjusting for anxiety/depression symptom scores (Figure 3B*i*).

3.5 Cognitive distancing and antidepressant use have opposite effects on the weighting of choices in subjective happiness

Next, we quantified the associations between treatments—cognitive distancing and antidepressant use—and choice-dependent parameters (i.e., w_2^p and w_3^p) which control the extent of trial-to-trial fluctuations in affect.

There was limited evidence for an association between either treatment and weights on recent prediction errors (w_3^p; Figure 3A*iv* & Figure 3B*iv*). However, there was some evidence that cognitive distancing lowered weighting of recent expected (choice) values in happiness ratings (Figure 3A*iii*), with 1.83% (95% HDI for multiplier=[0.961, 1.003]) lower

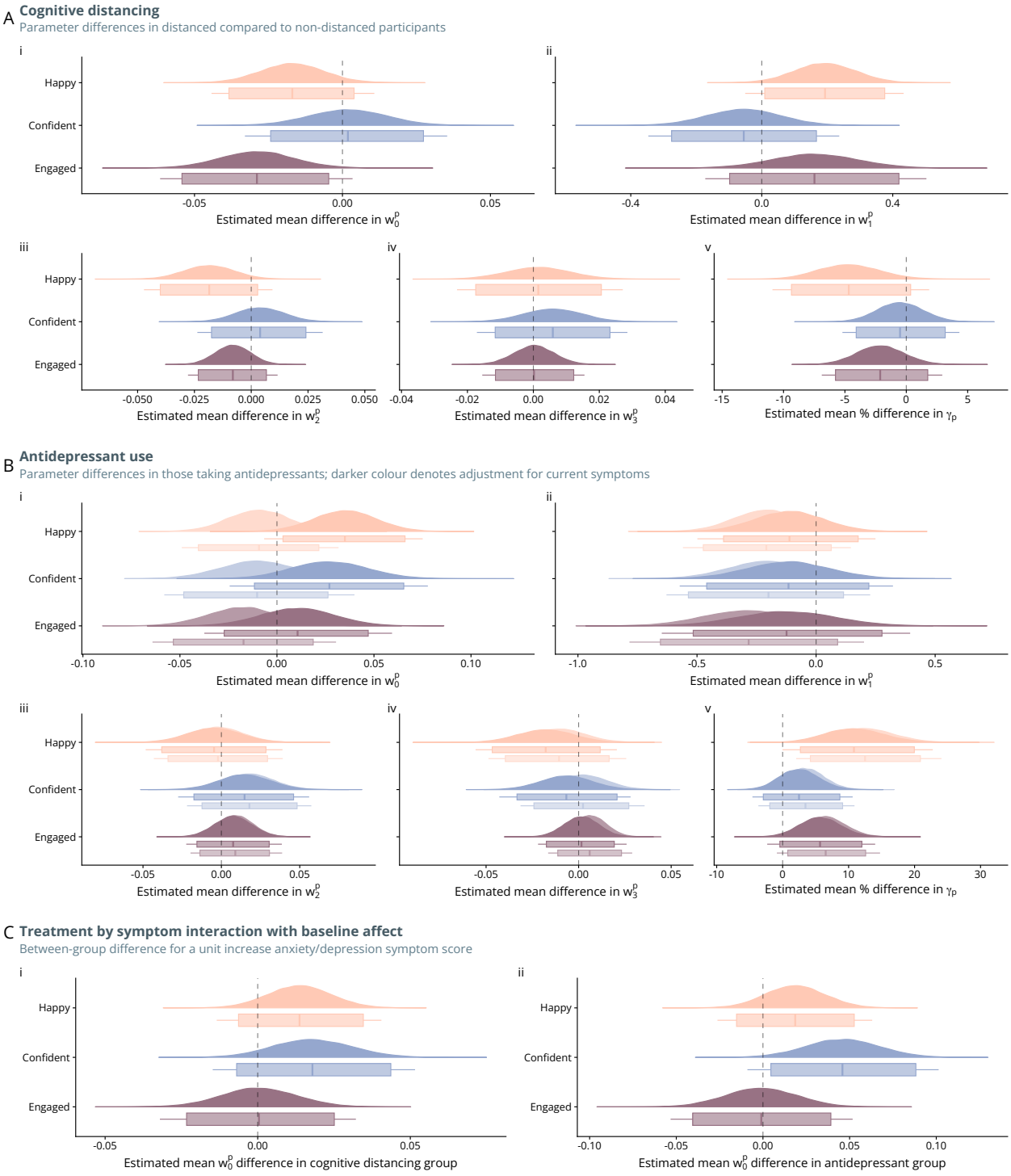

Figure 3: **Associations between treatments and affect model parameters. A-B**. Estimated mean differences in individuals' baseline affect (w_0^p), drift in affect over time (w_1^p) and forgetting rate of previous trials' expected values and prediction errors (γ_p), in participants practicing cognitive distancing (A) and taking antidepressants (B; darker colour denotes adjustment for current anxiety/depression symptoms). **C**. Interactions between cognitive distancing (*i*) and antidepressant use (*ii*) and higher anxiety/depression symptom scores, with respect to baseline affect. **D**. Associations between antidepressant use and expected value weights in affect rating computation: main effect (*i*), interaction with trial lag (*ii*), and posterior mean parameters for weighting of previous choices' expected values in engagement ratings ($w^p_{2_{(-t')}}$) by antidepressant group (*iii*). In all plots, boxplot boxes denote 95% HDIs, and lines denote 99% HDIs.

odds of increased happiness ratings for the same weighted sum of recent expected values in distanced participants. Meanwhile, exploratory analyses with an extended between-rating model showed a specific effect of antidepressant use on the weighting of recent expected values in engagement and confidence ratings (see Supplementary Results & Figure S4B*iii-iv*).

Current antidepressant use was associated with less forgetting of choices and outcomes in happiness ratings (12.5% higher γ_{happy}; 95% HDI for multiplier = [1.04, 1.21]) and engagement (6.52% higher $\gamma_{engaged}$; 95% HDI for multiplier = [1.01, 1.13]); Figure 3B*iii*), suggesting higher weighting of earlier trials' expected values and prediction errors in subsequent affect ratings. Evidence for this association remained, albeit slightly weakened, after additionally adjusting for current anxiety/depression symptoms (Figure 3B*iii*), which were themselves positively associated with γ_{happy} and (to a lesser extent) γ_{engaged} (Figure S3A*iii*). Notably, the converse effect was seen in distancing participants, with the psychological intervention associated with lower happiness forgetting factors, albeit with weak evidence (4.69% lower γ_{happy}; 95% HDI for multiplier = [0.906, 1.004]; Figure 3A*iii*).

3.6 Cognitive distancing and antidepressant use dampen negative associations between baseline affect and anxiety/depression symptoms

Lastly, we explored whether the negative associations between choice-independent affective dynamics and transdiagnostic anxiety/depression symptoms were altered by cognitive distancing or current antidepressant use, by including treatment by symptom interactions in outcome GLMs.

We found that both the distancing intervention and antidepressant use weakened the negative associations between baseline happiness and confidence, but not engagement. Specifically, distancing individuals with higher anxiety/depression scores had higher baseline happiness and confidence (mean 1.37-point higher w_0^{happy} and 1.79-point higher $w_0^{\text{confident}}$ per SD increase in anxiety/depression score respectively) relative to non-distancing participants with the same symptom scores, though with weak evidence (95% HDIs for this distancing by symptom interaction = [-0.63, 3.46] for baseline happiness and [-1.11, 6.12] for confidence (Figure 3C*i*). These effects were mirrored in participants taking antidepres-

sants. The evidence for an antidepressant by anxiety/depression symptom interaction with respect to baseline happiness was fairly weak (1.86-point higher w_0^{happy} per SD increase in anxiety/depression score; 95% HDI = [-1.52, 5.27]), but the evidence for the corresponding interaction effect on baseline confidence was stronger (mean 4.59-point higher $w_0^{\text{confident}}$ per SD increase in anxiety/depression score; 95% HDI = [0.559, 12.53]; Figure 3C*ii*).

4 Discussion

Here, we applied a computational model of momentary happiness which assumes fluctuations in affect ratings depend simply on baseline affect, its drift over time, and recency-decayed expected and received outcomes. By extending this model to also capture learning, we were able to link objective behaviour to subjective feelings across distinct affective states—happiness, confidence, and engagement ratings—and show that a a core component of psychological therapy, cognitive distancing, and antidepressant medication use have different effects on affective dynamics, but converge to alter affective biases associated with symptoms of mental ill-health.

There were distinct effects of both treatments on affective dynamics. Randomisation to a psychotherapeutic intervention, cognitive distancing, attenuated declines in happiness and engagement over time, adding to our previously reported findings that this psychotherapeutic technique alters aspects of reward learning[34]. Self-reported antidepressant use, meanwhile, was associated with higher baseline happiness and confidence after adjustment for current anxiety/depression symptoms (as antidepressant use was not randomised), which is consistent with evidence that antidepressants exert positive affective biases[28]. Subsequent exploration of changes in affect revealed further mechanistic divergence: current antidepressant use was associated with lower recency biases across all scales (i.e., forgetting factors closer to one), and cognitive distancing reduced the weighting of expectations and higher recency bias in happiness ratings. Together, these results suggest that psychiatric treatments act to alter the contribution of objective learning-related quantities to subjective value judgements.

Consistent with extensive evidence[14,18,19,26], we found negative associations between model-derived baseline affect (across all scales) and transdiagnostic psychiatric symptom measures derived from computational factor modelling[36,37]. This effect, which was strongest for anxiety/depression symptom scores, indicates a consistent, time-invariant negative affective bias which scales with mental ill-health symptom load. Critically, we found evidence for a convergent treatment by symptom interaction with baseline affect across both cognitive distancing and antidepressant use: negative associations between anxiety/depression symp-

toms and baseline happiness and confidence were attenuated in participants with higher anxiety/depression symptom scores. These results support the cognitive neuropsychological model of antidepressant action, whereby antidepressants are proposed to act acutely to revert negative or maladaptive affective biases[29,52], and suggest that cognitive distancing[31] and other components of psychotherapy may also act clinically to alter affective biases contributing to symptoms of mental ill-health. We propose that changes in affective dynamics should be investigated further in longitudinal studies as a computational predictor of subsequent symptom change.

Our methodological approach also extends previous work in two ways. First, we applied a theory-driven computational model which allows for fluctuations in momentary happiness as a function of the history of expected values and prediction errors resulting from those expectations, which has been primarily validated in tasks where learning is not required[5]. We not only show that this model can capture happiness ratings in a task where expected values are never explicitly available and have to be learned from experience, but can also accurately capture variation in subjective ratings of confidence and engagement. Second, we found evidence for drift in happiness over time, replicating recent work which characterised 'mood drift over time'[44], and extended this to both confidence and engagement. We also found that this drift was strongly associated with self-reported fatigue (Figure S5; see Supplementary Results for more details). We note that we did not find evidence of a previously reported effect—reduced mood drift over time with increased depressive symptoms[44]—instead finding evidence to the contrary (Figure 1A*ii*). However, this work primarily reported evidence from short gambling tasks[44]; our results indicate that associations between affective drift and mental ill-health symptoms are not task-invariant.

We note several limitations. Firstly, there were limitations in our outcome measures. Transdiagnostic measures of mental health psychopathology were estimated using questionnaire subsets[34,39], precluding investigation of associations between parameters and individual diagnostic scales. Antidepressant use, meanwhile, was self-reported, non-randomised, and we did not collect information concerning length of treatment. Secondly, our use of a single computational model for all three affect scales is a powerful approach, but limited in its ability to truly contrast trial-to-trial fluctuations in each individual rating scale, as

we are only comparing the contribution of a small number of computational components (i.e., affective weights) to a fraction of their variation; the residual (scale-specific) variation is likely also important in explaining how these ratings overlap and differ. Third, as model complexity meant only approximate (mean-field variational, rather than sampling-based) inference was viable, we were unable to account for uncertainty in estimates of individual-level posterior mean parameters in associations with quantities of interest (e.g., by using precision-weighted GLMs), as the posterior covariance matrix cannot accurately capture local interdependencies, meaning that parameter precisions are not reliable enough for uncertainty-weighted outcome models[45].

To conclude, we integrated objective choice behaviour in a learning task with trial-to-trial affect ratings across three distinct states—happiness, confidence, and engagement—within a unified computational model. This enabled us to uncover associations between model parameters and treatments for mental health conditions, offering new insights into their underlying mechanisms-of-action. Our results demonstrate the critical importance of affective biases in the maintenance and updating of affective states in mental ill-health, and indicate that existing, effective treatments can be understood at least in part as acting to shift these biases towards the healthy range.

Acknowledgements

This study was funded by an AXA Research Fund Fellowship awarded to C.L.N. (G102329) and the Medical Research Council (MC_UU_00030/12). C.L.N. is funded by a Wellcome Career Development Award (226490/Z/22/Z) and acknowledges support by the NIHR Cambridge NIHR Biomedical Research Centre (BRC-1215-20014). Q.D. is funded by a Wellcome Trust PhD studentship. Q.D. and Q.J.M.H acknowledge support by the NIHR UCLH BRC. Q.J.M.H. acknowledges grant funding from the NIHR, Wellcome Trust, Carigest S.A. and Koa Health. R.B.R. is supported by the National Institute of Mental Health (R01MH124110). R.B.R. holds equity in Maia.

Data and code availability

All code to replicate the analyses here can be found in accompanying Jupyter notebooks, alongside cleaned, anonymised data. See the GitHub repository for more details.

Rights retention

Acronyms

ADVI Automatic Differentiation Variational Inference
BCa Bias-Corrected and accelerated
CI Confidence Interval
ELPD Expected Log Pointwise Predictive Density
GLM Generalised Linear Model
HDI Highest Density Interval
HBREC Human Biology Research Ethics Committee
LOO Leave One Out
MSE Mean Squared Error
MCMC Markov Chain Monte Carlo
NHS National Health Service
PLS Partial Least Squares
RL Reinforcement Learning
SD Standard Deviation
SSRI Selective Serotonin Reuptake Inhibitor

References

1. Oswald, A. J. & Wu, S. Objective Confirmation of Subjective Measures of Human Well-Being: Evidence from the U.S.A. *Science* **327,** 576–79. DOI (2010).
2. Diener, E., Oishi, S. & Tay, L. Advances in Subjective Well-Being Research. *Nature Human Behaviour* **2,** 253–60. DOI (2018).
3. Steptoe, A., Deaton, A. & Stone, A. A. Subjective Wellbeing, Health, and Ageing. *The Lancet* **385,** 640–48. DOI (2015).
4. Suh, E., Diener, E. & Fujita, F. Events and Subjective Well-Being: Only Recent Events Matter. *Journal of Personality and Social Psychology* **70,** 1091–1102. DOI (1996).
5. Rutledge, R. B., Skandali, N., Dayan, P. & Dolan, R. J. A Computational and Neural Model of Momentary Subjective Well-Being. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 12252–57. DOI (2014).
6. Taquet, M., Quoidbach, J., Montjoye, Y.-A., Desseilles, M. & Gross, J. J. Hedonism and the Choice of Everyday Activities. *Proceedings of the National Academy of Sciences of the United States of America* **113,** 9769–73. DOI (2016).
7. Blain, B. & Rutledge, R. B. Momentary Subjective Well-Being Depends on Learning and Not Reward. *eLife* **9,** 1–27. DOI (2020).
8. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and Certainty: Distinct Probabilistic Quantities for Different Goals. *Nature Neuroscience* **19,** 366–74. DOI (2016).
9. Adler, W. T. & Ma, W. J. Comparing Bayesian and Non-Bayesian Accounts of Human Confidence Reports. *PLOS Computational Biology* **14,** 1006572. DOI (2018).
10. Navajas, J. *et al.* The Idiosyncratic Nature of Confidence. *Nature Human Behaviour* **1,** 810–18. DOI (2017).
11. Li, H.-H. & Ma, W. J. Confidence Reports in Decision-Making with Multiple Alternatives Violate the Bayesian Confidence Hypothesis. *Nature Communications* **11,** 2004. DOI (2020).
12. Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. & Rushworth, M. F. S. Weighing up the Benefits of Work: Behavioral and Neural Analyses of Effort-Related Decision Making. *Neural Networks* **19,** 1302–14. DOI (2006).
13. Ang, Y.-S., Gelda, S. E. & Pizzagalli, D. A. Cognitive Effort-Based Decision-Making in Major Depressive Disorder. *Psychological Medicine,* 1–8. DOI (2022).
14. Rutledge, R. B. *et al.* Association of Neural and Emotional Impacts of Reward Prediction Errors With Major Depression. *JAMA Psychiatry* **74,** 790–97. DOI (2017).
15. Barge-Schaapveld, D. Q., Nicolson, N. A., Berkhof, J. & deVries, M. Quality of Life in Depression: Daily Life Determinants and Variability. *Psychiatry Research* **88,** 173–89. DOI (1999).
16. Taquet, M., Quoidbach, J., Gross, J. J., Saunders, K. E. & Goodwin, G. M. Mood Homeostasis, Low Mood, and History of Depression in 2 Large Population Samples. *JAMA Psychiatry* **77,** 944–51. DOI (2020).
17. Rouault, M., Seow, T., Gillan, C. M. & Fleming, S. M. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry* **84,** 443–51. DOI (2018).
18. Hoven, M. *et al.* Abnormalities of Confidence in Psychiatry: An Overview and Future Perspectives. *Translational Psychiatry* **9,** 268. DOI (2019).
19. Hoven, M., Denys, D., Rouault, M., Luigjes, J. & Holst, R. How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach. *Nature Mental Health* **1,** 337–345. DOI (2022).

20. Katyal, S., Huys, Q. J., Dolan, R. J. & Fleming, S. M. Distorted learning from local metacognition supports transdiagnostic underconfidence. *Nature Communications* **16,** 1854. DOI (2025).
21. Cléry-Melin, M.-L. *et al.* Why Don't You Try Harder? An Investigation of Effort Production in Major Depression. *PloS One* **6,** 23178. DOI (2011).
22. Pessiglione, M., Vinckier, F., Bouret, S., Daunizeau, J. & Bouc, R. L. Why Not Try Harder? Computational Approach to Motivation Deficits in Neuro-Psychiatric Diseases. *Brain* **141,** 629–50. DOI (2018).
23. Husain, M. & Roiser, J. P. Neuroscience of Apathy and Anhedonia: A Transdiagnostic Approach. *Nature Reviews Neuroscience* **19,** 470–84. DOI (2018).
24. Disner, S. G., Beevers, C. G., Haigh, E. A. P. & Beck, A. T. Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience* **12,** 467–477. DOI (2011).
25. Lewis, G. *et al.* Variation in the recall of socially rewarding information and depressive symptom severity: a prospective cohort study. *Acta Psychiatrica Scandinavica* **135,** 489–498. DOI (2017).
26. Bylsma, L. M., Taylor-Clift, A. & Rottenberg, J. Emotional Reactivity to Daily Events in Major and Minor Depression. *Journal of Abnormal Psychology* **120,** 155–67. DOI (2011).
27. Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as Representation of Momentum. *Trends in Cognitive Sciences* **20,** 15–24. DOI (2016).
28. Harmer, C. J. Serotonin and Emotional Processing: Does It Help Explain Antidepressant Drug Action? *Neuropharmacology* **55,** 1023–28. DOI (2008).
29. Harmer, C. J., Duman, R. S. & Cowen, P. J. How Do Antidepressants Work? New Perspectives for Refining Future Treatment Approaches. *The Lancet Psychiatry* **4,** 409–18. DOI (2017).
30. Beck, A. T. Thinking and depression: II. Theory and therapy. *Archives of General Psychiatry* **10,** 561–571. DOI (1964).
31. Kross, E., Gard, D., Deldin, P., Clifton, J. & Ayduk, O. "Asking Why" from a Distance: Its Cognitive and Emotional Consequences for People with Major Depressive Disorder. *Journal of Abnormal Psychology* **121,** 559–69. DOI (2012).
32. Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science* **306,** 1940–43. DOI (2004).
33. Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T. & Hutchison, K. E. Genetic Triple Dissociation Reveals Multiple Roles for Dopamine in Reinforcement Learning. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 16311–16. DOI (2007).
34. Dercon, Q., Mehrhof, S. Z. *et al.* A Core Component of Psychological Therapy Causes Adaptive Changes in Computational Learning Mechanisms. *Psychological Medicine,* 1–11. DOI (2023).
35. Powers, J. P. & LaBar, K. S. Regulating emotion through distancing: A taxonomy, neurocognitive model, and supporting meta-analysis. *Neuroscience & Biobehavioral Reviews* **96,** 155–173. DOI (2019).
36. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a Psychiatric Symptom Dimension Related to Deficits in Goal-Directed Control. *eLife* **5** (ed Frank, M. J.) e11305. DOI (2016).
37. Wise, T., Robinson, O. & Gillan, C. Identifying Transdiagnostic Mechanisms in Mental Health Using Computational Factor Modeling. *Biological Psychiatry* **93,** 690–703. DOI (2022).
38. Palan, S. & Schitter, C. Prolific.Ac—A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* **17,** 22–27. DOI (2018).
39. Wise, T. & Dolan, R. J. Associations between Aversive Learning Processes and Transdiagnostic Psychiatric Symptoms in a General Population Sample. *Nature Communications* **11,** 4179. DOI (2020).

40. Kao, C.-H., Feng, G. W., Hur, J. K., Jarvis, H. & Rutledge, R. B. Computational Models of Subjective Feelings in Psychiatry. *Neuroscience & Biobehavioral Reviews* **145,** 105008. DOI (2023).
41. Forbes, L. & Bennett, D. The effect of reward prediction errors on subjective affect depends on outcome valence and decision context. *Emotion* **24,** 894–911. DOI (2024).
42. Ferrari, S. & Cribari-Neto, F. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics* **31,** 799–815. DOI (2004).
43. Smithson, M. & Verkuilen, J. A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. *Psychological Methods* **11,** 54–71. DOI (2006).
44. Jangraw, D. C. *et al.* A Highly Replicable Decline in Mood during Rest and Simple Tasks. *Nature Human Behaviour* **7,** 596–610. DOI (2023).
45. Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. *Automatic Differentiation Variational Inference* 2016. Preprint on *arXiv*.
46. Stan Development Team. *Stan Modelling Language Users Guide and Reference Manual* v2.31.0 (2022). LINK.
47. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing* **27,** 1413–32. DOI (2016).
48. Magnusson, M., Andersen, M. R., Jonasson, J. & Vehtari, A. *Leave-One-Out Cross-Validation for Bayesian Model Comparison in Large Data* 2020. Preprint on *arXiv*.
49. Greenland, S. *et al.* Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations. *European Journal of Epidemiology* **31,** 337–50. DOI (2016).
50. Goodrich, B., Gabry, J., Ali, I. & Brilleman, S. *rstanarm: Bayesian applied regression modeling via Stan.* (2020). LINK.
51. Wold, S., Ruhe, A., Wold, H. & Dunn III, W. J. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific and Statistical Computing* **5,** 735–43. DOI (1984).
52. Harmer, C. J., Goodwin, G. M. & Cowen, P. J. Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *British Journal of Psychiatry* **195,** 102–108. DOI (2009).
53. Bürkner, P.-C. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* **80,** 1–28. DOI (2017).

Supplementary Material

Supplementary Methods

Interpretation of model-derived parameters

In the Results, we report intercept (i.e., baseline affect) parameters (w_0^p) following an inverse logit transformation to allow interpretation of the GLM coefficient on the original (point) scale (from 0 to 1) as the difference in baseline affect rating between individuals differing only in the covariate of interest (by one unit). Other GLM-estimated weight parameter differences are interpreted as the difference in the (log) odds of an increase in affect rating between individuals differing only in the covariate that the parameter weights.

For intuition, consider an estimated GLM coefficient of 0.1 for the distancing group in relation to baseline happiness (w_0^{happy}) inverse logit transformed to a 0 to 1 scale. This result indicates an estimated 10-point (on the 0-100 scale) higher baseline happiness rating in distanced participants after covariate adjustment. Meanwhile, a GLM coefficient for the distancing group of 0.1 in relation to w_1^{happy} for the time elapsed (overall) model (i.e., where w_1^{happy} is the adjusted log odds ratio for an increase in happiness for a one hour increase in elapsed time) suggests that the distanced group were 10.5% more likely than non-distanced participants to have an increase in happiness after an hour (i.e., the estimated multiplier is $\exp(0.1) = 1.105$)[b]. Similarly, a GLM coefficient for the distancing group of 0.1 in relation to w_2^{happy} suggests that a unit increase in Q-value is 10.5% more likely to produce an increase in happiness rating, suggesting a higher contribution of (recent) expected values in the affect rating computation.

Between-rating RL-affect model

Model definition and fit

In an exploratory analysis, we modified (6) to partition the weights on expected values and prediction errors (w_2^p and w_3^p) into individual weights on the outcomes of each previous trial since the previous rating (in our task, this could be up to four intervening outcomes plus the current trial). As such, the lagged trial weight parameters $w_{2_{(-t')}}^p$ and $w_{3_{(-t')}}^p$ can be seen as

[b] Note the exact interpretation (i.e., greater increase vs. slower decline in odds) depends on the GLM intercept term, which in this case would be the estimated w_1^{happy} in non-distanced participants after covariate adjustment.

capturing the between-rating change in affective dynamics. We hence modify (6) as follows,

$$\log\left(\frac{\mu_t^{(i)}}{1-\mu_t^{(i)}}\right) = w_0^p + w_1^p\,\tau_t + \sum_{t'=0}^{I} w^p_{2_{(-t')}}\,Q_{t-t'}(s_{t-t'},a_{t-t'}) + \sum_{t'=0}^{I} w^p_{3_{(-t')}}\,[r_{t-t'} - Q_{t-t'}(s_{t-t'},a_{t-t'})], \tag{7}$$

where I denotes the number of ratings between the current rating and the last rating of the same type; in this task, $I \leq 4$. $w^p_{k_{(-t')}}$ are the weights on the outcomes t' trials back, so $w^p_{k_{(0)}}$ is the weight on the outcome of the current trial, $w^p_{k_{(-1)}}$ is the weight on the outcome of the previous trial, and so on. This model was fit to choices and ratings on all three affect scales simultaneously in a hierarchical Bayesian manner via ADVI[45] as explained in Methods. The only difference was that lagged trial weights were assumed drawn from w_2^p and w_3^p priors specific to each individual which in turn were drawn from overall group-level hyperpriors (i.e., a three-level hierarchy).

Associations between between-rating RL-affect model parameters and treatments

The associations between the weights on expected values and prediction errors for individual trials and cognitive distancing and antidepressant use were quantified using multilevel Bayesian linear regression models implemented in the `brms` R package[53] and CmdStan[46]. All models controlled for the same covariates (i.e., age, gender, digit span), but also included a participant-level random intercept and slopes (on trial lag) to account for the fact there were five parameters per person for each affect rating, as well as the main effect of trial lag and its interaction with each treatment. Separate models were fit for each affect rating and parameter (i.e., $w_{2_{(-t')}}$ and $w_{3_{(-t')}}$).

Parameter recovery

Joint RL-affect model with mood drift over time

To test whether we could recover known parameter values from the best-fitting model (i.e., dual learning rate, time elapsed over time), we simulated one hundred datasets (including choices and affect ratings) with parameters drawn from the following distributions:

We then fit the model to these simulated data with approximate inference, and compared the posterior mean parameter estimated to those known to have generated the data. We

Parameter	Distribution
α_{reward}, α_{loss}	$\text{Beta}(1.5, 3)$
$0.1\beta^{\dagger}$	$\text{Beta}(3, 4)$
w_0^p	$\text{Normal}(0, 0.5)$
w_1^p	$\text{Normal}(-0.5, 1)$
w_2^p, w_3^p	$\text{Normal}(0.2, 0.1)$
γ_p	$\text{Beta}(2, 2)$

Table S1: Parameter distributions used to simulate data and test parameter recovery. †i.e., so $\beta \in [0, 10]$

found that all parameters could be recovered with high accuracy (r >0.87; Figure S1A).

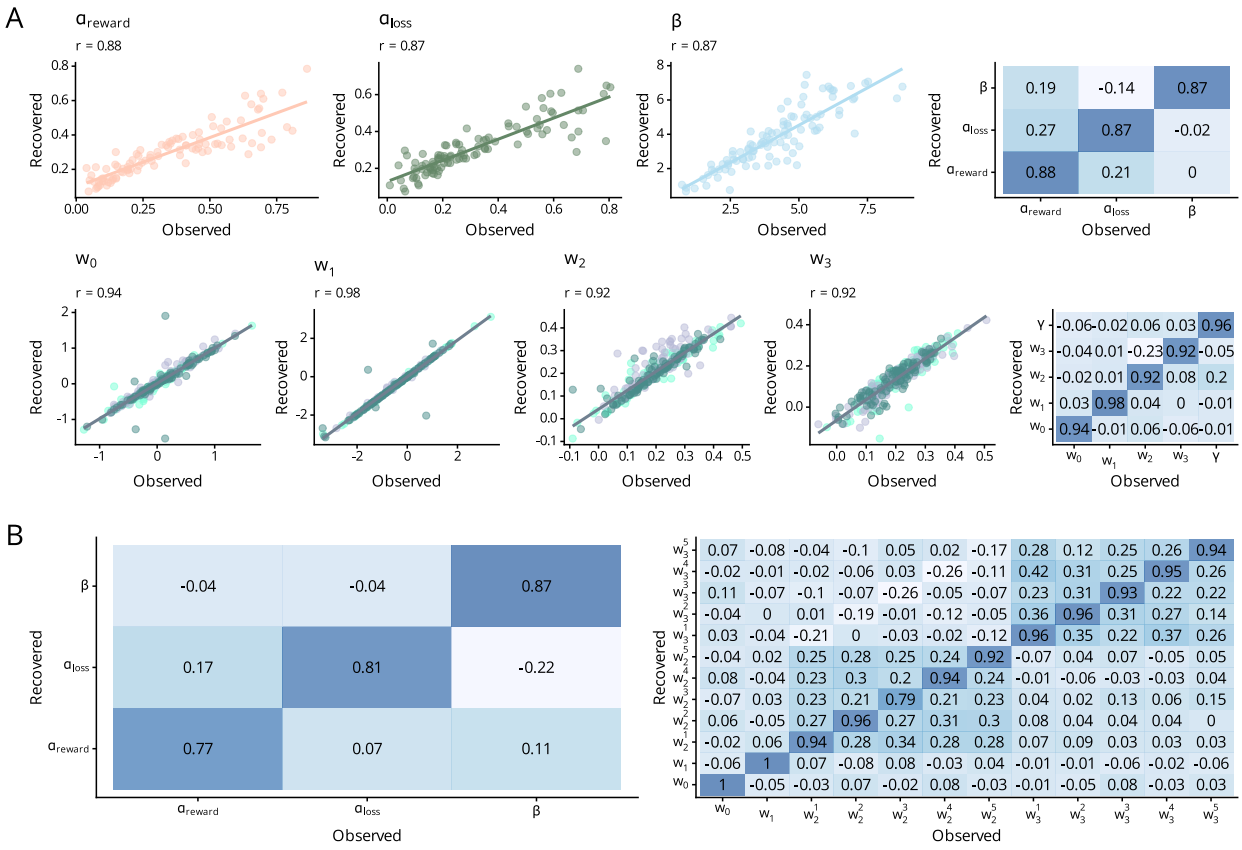

Figure S1: Parameter recovery for A) the joint RL-affect model, and B) the between-rating RL-affect model.

Between-rating RL-affect model

Parameter recovery for the between-rating RL-affect model was tested similarly, with one hundred simulated datasets. The parameter settings were identical to the above except for the time-dependent parameters (and the absence of γ_p in the model). Specifically, $w^p_{2_{(-t')}}$ and

$w^p_{3_{(-t')}}$ were sampled from $Beta(1, j)$ distributions, where t' is the trial lag, so that weights for earlier trials were weighted (on average) lower, as we see in the real dataset. Again, even with the high complexity of the model, we found that all parameters could be recovered with reasonably high accuracy (r >0.77; Figure S1B).

Comparison between results from models fit to choices alone (with sampling-based inference) vs. in the joint RL-affect model (with variational inference)

We previously reported results in this dataset where we fit Q-learning models to choices alone and compared parameters in distanced and non-distanced participants[34].

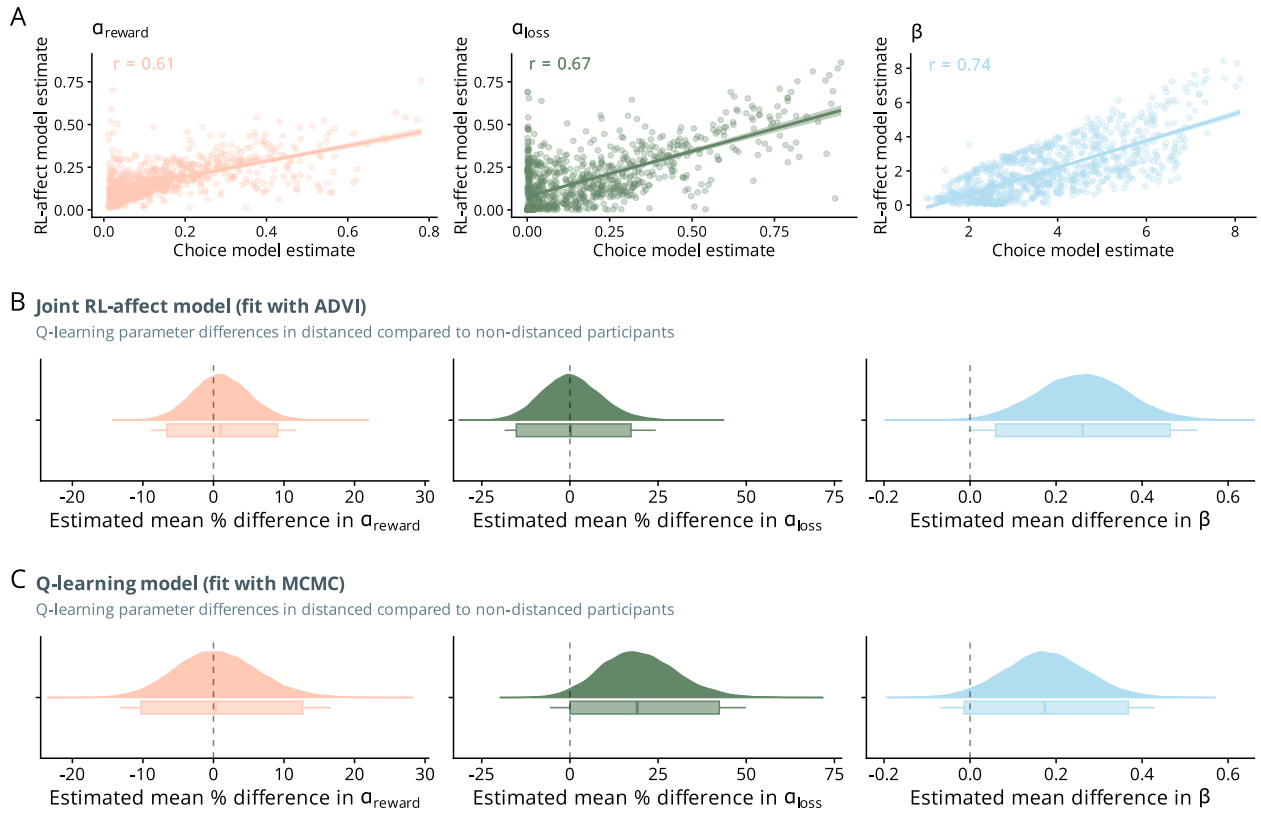

Figure S2: Comparison of Q-learning parameters and effects of distancing between dual learning rate models fit to choices alone and the joint RL-affect model additionally fit to affect ratings.

Besides the obvious difference—that the models were fit to choices alone as opposed to both choices and affect ratings—the models in our previous work[34] were fit to data via sampling-based inference (MCMC), as opposed to variational inference (ADVI[47]). However, despite these differences, we find that individuals' posterior mean Q-learning parameters from (*i*) dual learning rate models fit to choices alone with MCMC, and (*ii*) the best-fitting

joint RL-affect model fit choices and affect ratings simultaneously with ADVI are highly correlated with those we previously reported from Q-learning models fit to choices alone[34] in the same sample (α_{reward}: r=0.61 [95% CI = (0.57, 0.65)]; α_{loss}: r=0.67 [95% CI: 0.63, 0.70]; β: r = 0.74 [95% CI = (0.70, 0.76)]; Figure S2A). We also replicate a key result from our earlier work: higher inverse temperatures (β) in the distancing group (Figure S2B-C).

Supplementary Results

Group-level parameter estimates for the best-fitting joint RL-affect model

At the group-level, model-predicted baseline affect (w_0^p) was highest for engagement (group-level mean = 65.9 points; 95% HDI = [65.9, 66.0]), followed by happiness (group-level mean = 56.5, 95% HDI = [56.4, 56.6]), and lowest for confidence (group-level mean = 32.2, 95% HDI = [32.1, 32.2]). Affect drift over time (w_0^p) was steepest for engagement: an estimated 81.2% lower odds of increased engagement per hour (all other terms being equal; 95% HDI for multiplier = [0.187, 0.189]), compared to 54.1% lower confidence (95% HDI for multiplier = [0.457, 0.462]) and 37.7% lower happiness (95% HDI for multiplier = [0.620, 0.625]) per hour. Group-level means for weights on recent expected values (w_2^p) and prediction errors (w_3^p) were positive for all three ratings, showing higher affect with increased recent rewards, and were lowest for engagement (posterior mean [95% HDI] for multiplier: w_2^{happy} = 1.200 [1.198, 1.201], $w_2^{\text{confident}}$ = 1.216 [1.216, 1.218], w_2^{engaged} = 1.116 [1.114, 1.116]; w_3^{happy} = 1.136 [1.134, 1.136], $w_3^{\text{confident}}$ = 1.208 [1.207, 1.210], w_3^{engaged} = 1.109 [1.108, 1.109]). Lastly, the average decay factor was highest for confidence (group-level mean = 0.563; 95% HDI = [0.561, 0.565]), and lowest for happiness (group-level mean=0.407; 95% HDI = [0.405, 0.409]), suggesting higher weighting of outcomes of earlier trials (i.e., prior to the most recent trial) in confidence judgements compared to happiness ratings.

Compulsive behaviour and social withdrawal factor scores are also associated with altered affective dynamics

In addition to differences in baseline affect and its drift over time detailed in the main text, there was evidence that participants with higher compulsive behaviour scores placed more weight on recent expected values (higher w_2^p; 95% HDI excluding zero for w_2^{happy} only) and prediction errors (higher w_3^p) in their subjective affect judgements (Figure S3*i-iii*). The consequences of this were evident in the observed data. For example, there was evidence of a weak positive correlation between happiness rating variability and compulsive behaviour factor scores (correlation coefficient [95% CI] $r = 0.15$ [0.086, 0.212], $p < 0.0001$; Figure S3D), which can also be qualitatively observed by comparing mean-centred happiness ratings for participants in the bottom quartile versus the upper quartile of compulsive behaviour factor

scores (Figure S3E). That said, the actual effect of the estimated differences in w_2^{happy} and w_3^{happy} on ratings is small, as shown by simulating happiness ratings for individuals who differ only in having the estimated w_2^{happy} and w_3^{happy} for those with the 25$^{\text{th}}$ percentile versus the 75$^{\text{th}}$ percentile compulsive behaviour factor score (Figure S3E).

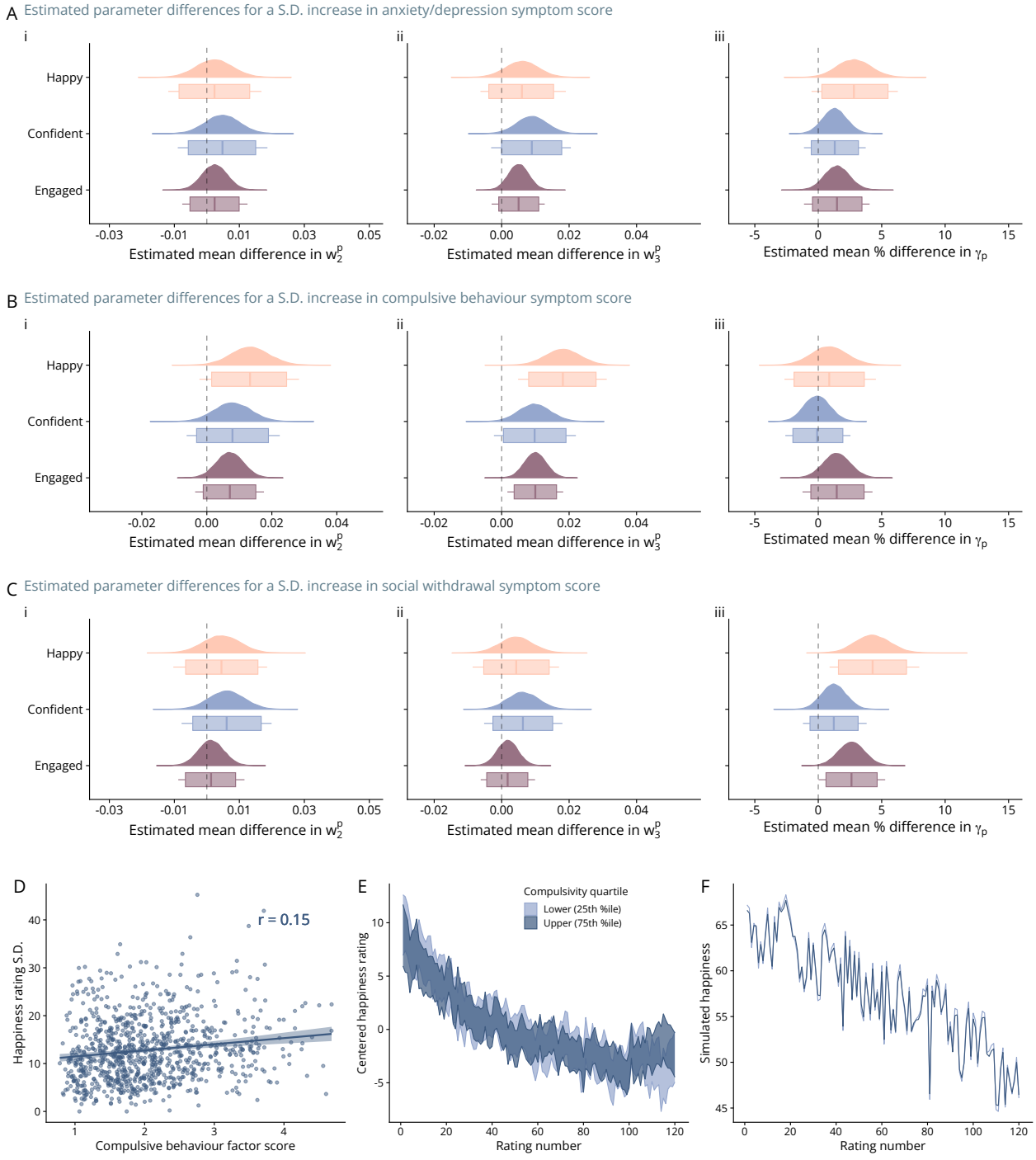

Figure S3: Associations between higher transdiagnostic psychiatric symptom factor scores and additional affect parameters (**A-C**), correlation between variance in happiness rating and compulsive behaviour score (**D**), and the simulated effect on happiness ratings of higher w_2^{happy} and w_3^{happy} (**E-F**).

We additionally found some weak evidence that increases in anxiety/depression factor were associated with slightly higher weighting of previous trials' expected values and prediction errors for happiness (γ_{happy}; e.g., trial-before-last weighted an estimated 4.04% higher; 95% HDI for multiplier = [1.003, 1.079]; Figure S3A*iii*). There was also some stronger evidence for a positive association between social withdrawal factor score and decay factors for happiness and engagement (Figure S3C*iii*), suggesting marginally higher weighting of previous trials' expected values and prediction errors in the computation of affect ratings in those with higher levels social withdrawal symptoms.

Antidepressant use is associated with increased weighting of previous choices' expected values in affect ratings

To further unpick the effects of treatments on the weighting of previous outcomes, we fit a more flexible between-rating RL-affect model (equation 7). This model, which allowed for different weights on previous expected values ($w^p_{2_{(-t')}}$) and prediction errors ($w^p_{3_{(-t')}}$) since the previous rating (up to five trials back), was able to capture the ratings well, albeit with marginally worse accuracy than the winning drift over time model (mean [SD] pseudo-R^2 for between-rating RL-affect model = 0.39 [0.22-0.25] across all three ratings). We then related $w^p_{2_{(-t')}}$ and $w^p_{3_{(-t')}}$ from each rating type separately to both treatments via multilevel GLMs with participant-level random intercepts and slopes (on trial lag), adjusting for age, gender, and digit span as before.

Parameters from this between-rating model suggested limited evidence for a difference between distancing and non-distancing participants in the weighting of the most recent or intervening outcomes in their affective judgements (Figure S4A). There was also no evidence of an effect of either treatment on weightings of prediction errors from previous trials (Figure S4Aiii-iv & Figure S4Biii-iv). There was, however, some evidence of a small effect of antidepressant use on between-rating changes in affect: higher weighting of the most recent expected value in subjective affect ratings (Figure S4B*i*). The evidence for this was strongest for engagement, with a unit increase in the most recent Q-value associated with 4.33% higher odds of an increase in engagement rating (95% HDI for multiplier = [1.001, 1.089]). Furthermore, there was limited evidence of an accompanying interaction effect

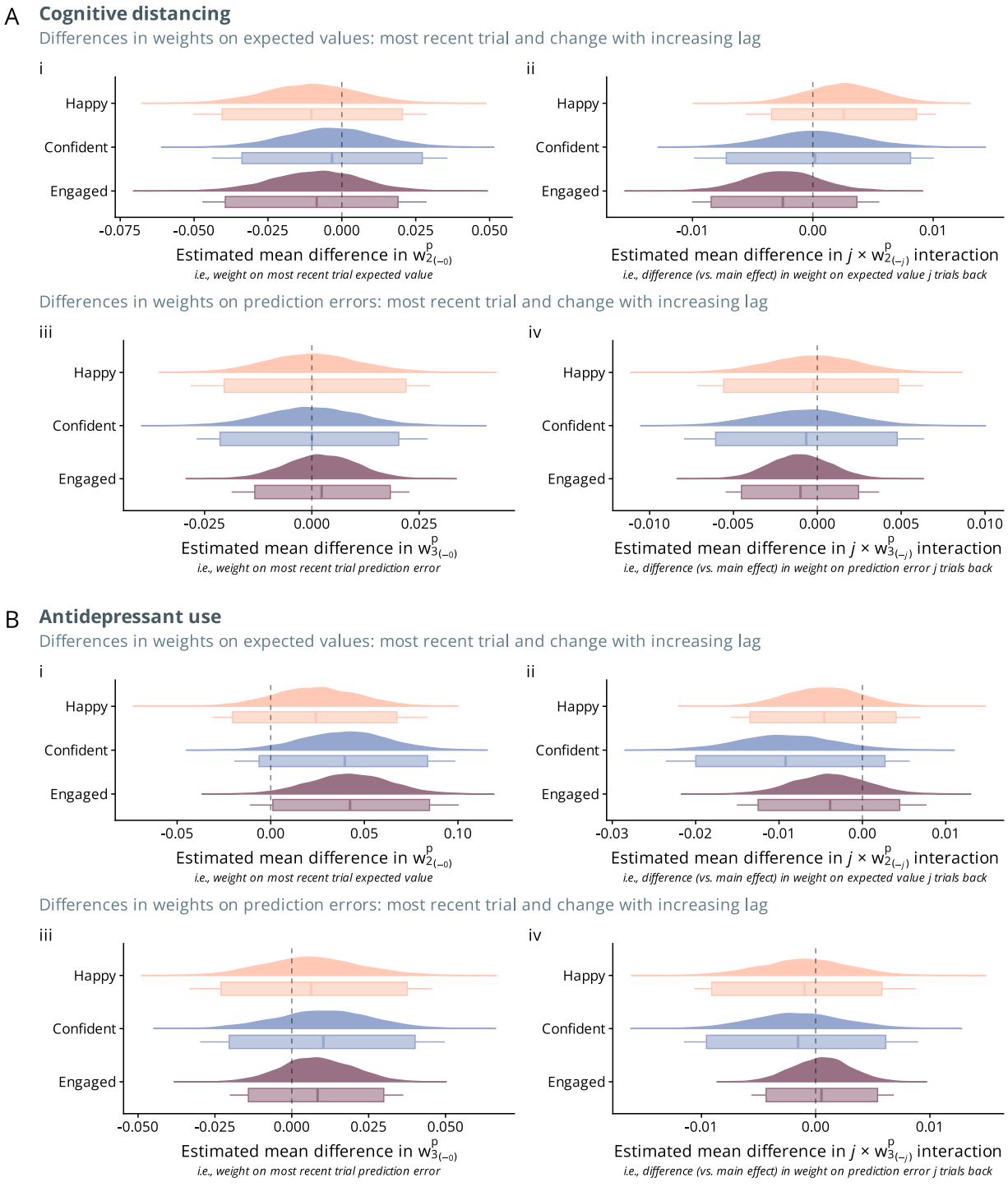

Figure S4: Effects of cognitive distancing (**A**) and antidepressant use (**B**) on expected value and prediction error parameters, derived from the between-rating RL-affect model.

(Figure S4B*ii*), suggesting the contribution of less recent expected values to engagement ratings was also marginally higher in participants taking antidepressants, which may in turn explain the higher forgetting factor γ_{engaged} (Figure 3B*v*).

Affective drift over time is associated with self-reported fatigue

Previous work on 'mood drift over time' has suggested it is mostly distinct from boredom and mind wandering[44]. Here, were able to test an additional aspect of this phenomenon, namely its relation to fatigue, as we asked participants the following question after the end of each of the six blocks: "How fatigued do you feel compared to the beginning of the block?". We hence examined the association between w_1^p and both participants' overall mean post-block fatigue and their change in post-block fatigue ratings over the course of the six training blocks. To quantify change in post-block fatigue, the six ratings were regressed on block number for each participant, with the regression coefficient ($\beta_{\Delta fatigue}$) on block number taken as the quantity of interest (i.e., higher values suggest increases in post-block fatigue ratings over time).

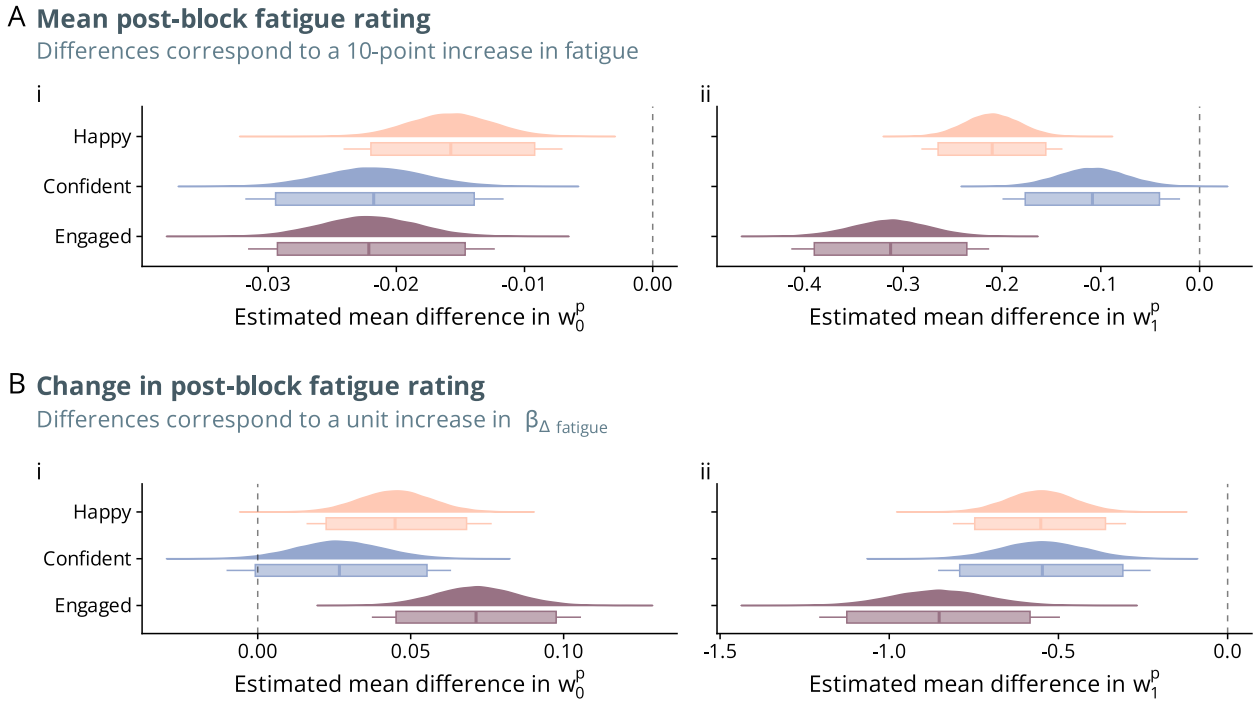

Figure S5: Associations between baseline affect and affective drift, and self-reported fatigue.

We found strong evidence, after adjusting for age, gender, digit span, and distancing group, that higher mean post-block fatigue ratings were associated with lower baseline affect (lower w_0^p; Figure S5A*i*) and decreased odds of higher affect ratings across the task (lower w_1^p; Figure S5A*ii*), across all three affect rating types (e.g., estimated mean 18.9% lower odds of increased happiness across the task with a ten-point increase in mean post-block fa-

tigue; 95% HDI for multiplier = [0.767, 0.856]). In addition, regression coefficients capturing change in post-block fatigue were strongly negatively associated with w_1^p across all rating types after adjusting for mean post-block fatigue, most strongly for engagement (estimated mean 5.74% lower odds of increase in engagement for a one-point per block in fatigue rating rate-of-change, 95% HDI for multiplier = [0.944, 0.968]; Figure S5B*ii*). Notably, associations in the opposite (positive) direction were observed between $\beta_{\Delta fatigue}$ and baseline affect, again most strongly for engagement (estimated 0.714-point increase in engagement rating for a one-point per block increase in fatigue rating rate-of-change; 95% HDI = [0.452, 0.975]; Figure S5B*i*). Speculatively, this may represent an effect of motivation, where participants who were more engaged towards the beginning of the task also exerted more effort, resulting in larger overall increases in fatigue and decreases in engagement.