nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0455-8

Behavioural and neural evidence for self-reinforcing expectancy effects on pain

Marieke Jepma[1,2*], Leonie Koban[2], Johnny van Doorn[1], Matt Jones[2] and Tor D. Wager[2]

Beliefs and expectations often persist despite evidence to the contrary. Here we examine two potential mechanisms underlying such 'self-reinforcing' expectancy effects in the pain domain: modulation of perception and biased learning. In two experiments, cues previously associated with symbolic representations of high or low temperatures preceded painful heat. We examined trial-to-trial dynamics in participants' expected pain, reported pain and brain activity. Subjective and neural pain responses assimilated towards cue-based expectations, and pain responses in turn predicted subsequent expectations, creating a positive dynamic feedback loop. Furthermore, we found evidence for a confirmation bias in learning: higher- and lower-than-expected pain triggered greater expectation updating for high- and low-pain cues, respectively. Individual differences in this bias were reflected in the updating of pain-anticipatory brain activity. Computational modelling provided converging evidence that expectations influence both perception and learning. Together, perceptual assimilation and biased learning promote self-reinforcing expectations, helping to explain why beliefs can be resistant to change.

Our past experiences drive our expectations about the future. This principle is a fundamental tenet underlying learning theory[1,2]. At the same time, our expectations can strongly influence how we experience events. This principle underlies decades of work on placebo effects[3–9], top-down influences on perception[10–13] and predictive coding[14–19]. The bidirectional interaction between expectations and experience can result in self-reinforcing phenomena—so-called 'self-fulfilling prophecies'—in many areas of human endeavour, including placebo and nocebo effects in medicine, stereotype effects on performance and behaviour, and economic growth and recession[20].

One domain in which self-reinforcing expectancy effects may be particularly powerful, and have important clinical implications, is pain perception. Previous studies have found that expectations about pain intensity—induced by previous experiences and instructions—result in the adjustment of pain responses toward the expected pain level[5,16,21–23]. Moreover, several studies have shown that expectancy effects on pain persist or even grow over time, in the absence of confirming evidence[24–31]. Such self-reinforcing expectancy effects are inconsistent with conventional reinforcement-learning principles[1,32]. In standard models of reinforcement learning, discrepancies between expected and actual outcomes—or 'prediction errors'—trigger expectation updating, such that expectations that are not confirmed by experience will extinguish.

The behavioural and brain mechanisms underlying resistance to extinction are largely unknown, and previous studies have not empirically demonstrated reciprocal, positive associations between expectations and experience. Thus, the idea of 'self-fulfilling prophecies' in brain-behaviour systems remains a theoretical conjecture. Here, we address this question in a laboratory setting, examining trial-to-trial dynamics in behaviour and functional magnetic resonance imaging (fMRI) activity related to expectations and pain. In two studies, we independently manipulated predictive cues and painful stimulus intensity, which allowed us to decouple the bidirectional influences of expectation and pain on each other.

Using this platform, we examine two non-mutually-exclusive ways in which expectations about pain can be self-reinforcing. First, expectations may modify the perceptual processing of nociceptive input, such that people actually feel what they expect. Findings that placebo and nocebo manipulations—involving suggestions of decreased and increased symptoms, respectively—influence pain-related activation in the spinal cord provide evidence that expectations can modify pain processing at a very early stage[33,34]. Similar modulation of perceptual processing may account for expectancy effects on appetitive experiences[35]. If the assimilation of sensory input towards expectations occurs at a processing stage prior to prediction-error computation, prediction errors will be diminished and hence expectation updating will be impeded.

A second possible mechanism underlying self-reinforcing expectancy effects is that expectations may bias experience-based learning. Specifically, people may update their expectations more when new evidence confirms, compared to when it does not confirm, their initial beliefs. Consistent with this idea, prior information about reinforcement probabilities can bias choices and suppress learning-related brain activation in probabilistic reward-learning tasks[36–40]. Expectations may induce biases in both (1) evaluation, that is, the reinforcement value assigned to outcomes[38,41], and (2) learning, that is, the degree to which new outcomes trigger expectation updating[37].

In the present study, we provide evidence that cue-based expectations influence both pain perception and learning. Higher pain expectations predicted larger subjective and neural pain responses, and larger pain responses in turn predicted higher subsequent expectations, generating a positive feedback loop between expectation and pain. Additionally, expectation updating for high-pain cues was strongest following higher-than-expected pain, while updating for low-pain cues was strongest following lower-than-expected pain, consistent with a confirmation bias in learning. Together, these effects promote persistent effects of initial expectations on pain in the face of predominantly non-confirming evidence.

[1]Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands. [2]Department of Psychology and Neuroscience and Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA. *e-mail: m.jepma@uva.nl

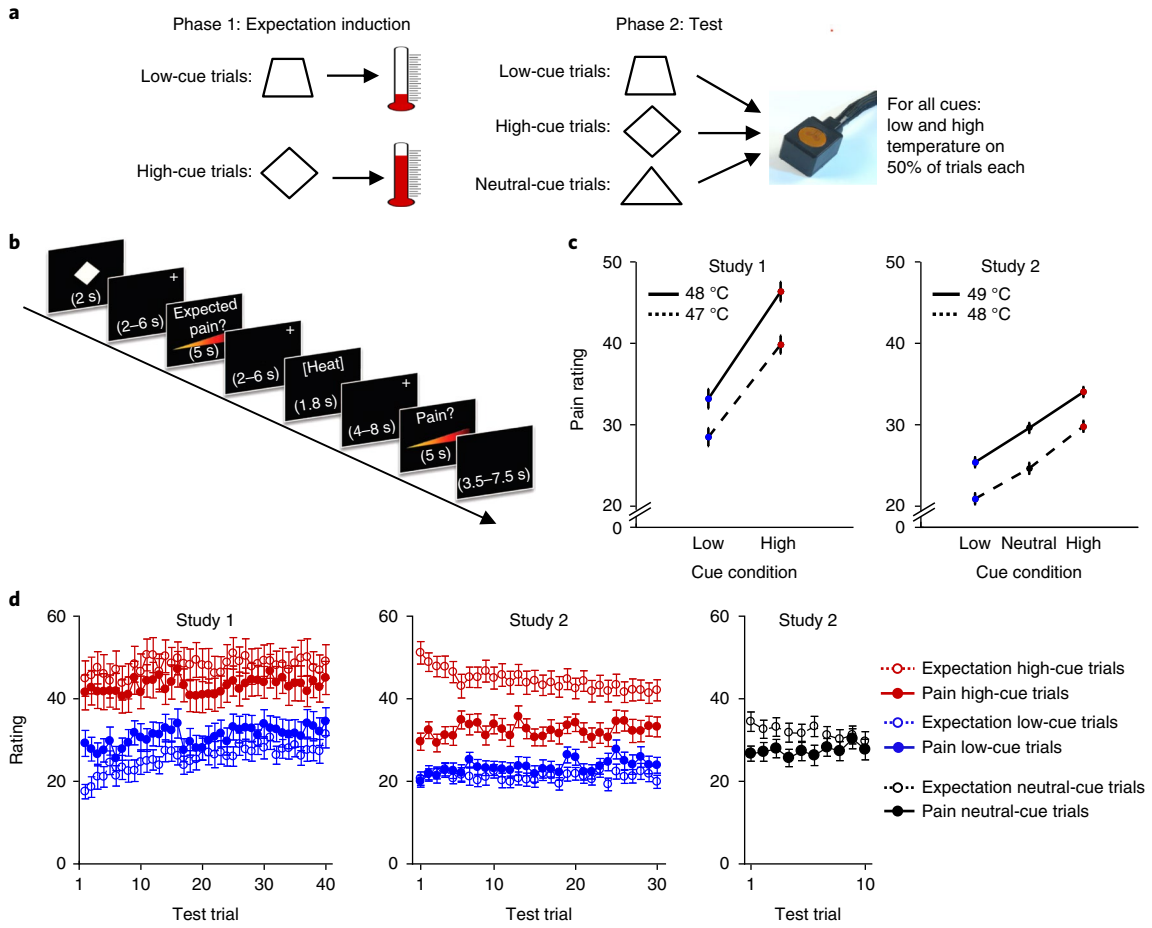

Fig. 1 | Experimental design and behavioural results. a, Cue-outcome pairings in the conditioning and test phase. Ten neutral-cue trials were included in Study 2 only. **b**, One test-phase trial in Study 2. Study 1 had slightly shorter inter-stimulus intervals. **c**, Pain rating as a function of stimulus temperature and cue type. Heat was applied to the inner forearm in Study 1 and to the lower leg (which is less sensitive) in Study 2, which explains the overall lower pain ratings in Study 2. Error bars indicate within-subject standard errors. Plots for Study 1 and Study 2 are based on data from 28 and 34 participants, respectively. **d**, Average expected (open circles) and experienced (filled circles) pain ratings as a function of cue type and trial. The difference between red and blue filled circles is the effect of cue type on pain ratings, which was robust in both studies, and did not disappear over time. The difference between red and blue open circles is the effect of cue type on pain expectations, which remained stronger than the effect on pain ratings throughout the test phase. Error bars indicate between-subject standard errors. Plots for Study 1 and Study 2 are based on data from 28 and 33 participants, respectively (one participant in Study 2 misunderstood the expected-pain rating procedure, and was excluded from all analyses and figures involving pain expectations).

Results

In two studies ($N=28$ and $N=34$), participants performed a learning procedure followed by a test phase[27]. During learning, participants viewed abstract visual cues paired with symbolic representations of heat, that is, pictures of thermometers (Fig. 1a). Some cues (low-pain cues) were consistently followed by low-temperature pictures (25–51% of the thermometer scale), and other cues (high-pain cues) were consistently followed by high-temperature pictures (73–93% of the thermometer scale). The purpose of this procedure was to create learned, conceptual associations between cues and heat intensity. In the subsequent test phase, both types of cues were repeatedly followed by noxious contact heat stimuli applied to participants' inner forearm (47–48 °C; Study 1) or lower leg (48–49 °C; Study 2). Importantly, unbeknownst to the participants, heat intensities during the test phase were matched for all cues (Fig. 1a), allowing a test of the causal effects of the cues on pain. Participants rated how much pain they expected following each cue, and how much pain they experienced following each heat stimulus (Fig. 1b). We focused our analyses on the test phase, including pain ratings in both studies and fMRI activity in Study 2. We have previously reported the cue effects on pain ratings and skin-conductance responses in Study 1[27]. Other results, including those on learning dynamics and computational modeling, were not included in previous reports.

We analysed the behavioural and neurologic pain signature (NPS) results using multi-level regression analyses on the single-trial data (also see Methods). Unless otherwise stated, the reported *t*-tests are tests of the distribution of the first-level regression coefficients against 0, and the reported confidence intervals (CI) are the 95% confidence intervals of the regression coefficients. We used bootstrapping for significance testing, which does not require the assumption of normality for valid inference.

Cue effects on pain ratings. As expected, participants' pain ratings increased with increasing temperature ($t(27)=9.4$, bootstrap $P<0.001$, Cohen's $d=1.9$, CI = 2.2 to 3.2 and $t(33)=9.4$, bootstrap $P<0.001$, $d=1.9$, CI = 3.5 to 5.1 in Study 1 and Study 2, respectively). Importantly, pain ratings were higher following high-pain cues than low-pain cues (Fig. 1c; $t(27)=7.6$, bootstrap $P<0.001$, $d=1.5$, CI = 4.6 to 7.5 and $t(33)=8.9$, bootstrap $P<0.001$, $d=1.6$,

CI = 3.4 to 5.2 in Study 1 and Study 2, respectively). In Study 1, the effect of cue type on pain rating was stronger for more intense heat, reflected in an interaction between cue type and temperature ($t(27) = 3.2$, bootstrap $P < 0.001$, $d = 0.91$, CI = 0.17 to 0.41). In Study 2, the effect of cue type did not differ between the two temperatures ($t(33) = 0.54$, bootstrap $P = 0.50$, CI = −0.48 to 0.26). In Study 2, we also included test trials with a novel, 'neutral' cue that had not been presented during the learning phase. Pain ratings on neutral-cue trials fell between those for low- and high-pain cues, and pain ratings for all three cue types differed significantly from each other, for both temperature levels (all bootstrap P values < 0.001; Fig. 1c).

The effect of cue type on pain rating was stable across test trials even though the cues no longer predicted heat intensity, providing initial evidence for potential 'self-reinforcing' effects (Fig. 1d). There was a trend towards a negative interaction between cue type and time (trial number) during the test phase ($t(27) = 1.8$, bootstrap $P = 0.07$, $d = 0.38$, CI = −0.04 to −0.0004 and $t(33) = 2.1$, bootstrap $P = 0.05$, $d = 0.34$, CI = −0.04 to −0.0004 in Study 1 and Study 2, respectively), reflecting a slight decrease in the cues' effects on pain over time. However, the effect of cue type was still large and highly significant at the end of the test phase (pain rating on the last high-pain-cue versus the last low-pain-cue trial: $t(27) = 4.6$, $P < 0.001$, $d = 0.87$ and $t(33) = 5.9$, $P < 0.001$, $d = 1.0$ in Study 1 and Study 2, respectively).

Cue effects on pain expectations. Analyses of expected-pain ratings showed that, in both studies, participants expected higher pain following high-pain cues than low-pain cues (Fig. 1d; $t(27) = 10.0$, bootstrap $P < 0.001$, $d = 1.9$, CI = 8.9 to 13.1 and $t(32) = 13.5$, bootstrap $P < 0.001$, $d = 2.5$, CI = 10.0 to 13.2 in Study 1 and Study 2, respectively). Furthermore, pain expectations for high-pain cues were consistently worse than experience (t-tests on average expected minus average experienced pain, $t(27) = 7.5$, $P < 0.001$, $d = 1.4$, CI = 3.5 to 6.2 and $t(32) = 7.1$, $P < 0.001$, $d = 1.2$, CI = 8.6 to 15.5 in Study 1 and Study 2, respectively), whereas pain expectations for low-pain cues were consistently better than experience (t-tests on average expected minus average experienced pain, $t(27) = 4.5$, $P < 0.001$, $d = 0.85$, CI = −7.5 to −2.8 and $t(32) = 3.1$, $P = 0.005$, $d = 0.54$, CI = −3.8 to −0.77 in Study 1 and Study 2, respectively), indicating a lack of extinction. Pain expectations for the neutral cues fell between those for low- and high-pain cues. In contrast to the low- and high-pain-cue trials, pain and expected-pain ratings on neutral-cue trials converged over time (Fig. 1d).

Cue effects on pain-related brain activity. We previously reported that, like pain ratings, heat-evoked skin conductance responses were also larger following high-pain cues than low-pain cues[27]. In Study 2, we examined the effects of the cues on pain-related brain activity. Heat-evoked activity in several areas—including the anterior midcingulate cortex, insula, thalamus and parts of the midbrain—was stronger on high- than low-pain-cue trials (Fig. 2a). These brain areas have been related to various aspects of pain processing[42–45]. For example, electrical stimulation of the insula can produce pain in humans[46], and lesions and pharmacological inactivation of the anterior cingulate cortex disrupt pain-avoidance behaviours in animals[47,48]. However, all these brain areas have also been associated with cognitive and emotional functions that are unrelated to pain; hence they are not specific to pain.

To address this issue, we recently developed a multivariate pattern of fMRI activity found to be sensitive and specific to physical pain in multiple previous studies[49–52]: the neurologic pain signature (NPS)[53]. We computed the strength of expression of the NPS by calculating the dot product of the NPS pattern weights and the activation map for each heat application period (obtained from the subject-level fMRI analysis; see Methods). This resulted in one scalar value—the NPS response—per trial for each participant.

As expected, higher NPS responses predicted higher pain ratings within participants, controlling for temperature and cue type ($t(33) = 3.1$, bootstrap $P < 0.001$, $d = 0.53$, CI = 0.01 to 0.03; Fig. 2b), and NPS responses were stronger for 49 °C than for 48 °C heat ($t(33) = 3.9$, bootstrap $P < 0.001$, $d = 0.77$, CI = 2.8 to 7.1; Fig. 2c). Importantly, NPS responses were larger on trials with high-pain than low-pain cues (main effect of cue type, $t(33) = 3.2$, bootstrap $P < 0.001$, $d = 0.63$, CI = 1.3 to 4.4; Fig. 2c), indicating that the cues modulated the nociceptive and/or pain-generation processes that drive the NPS response. Furthermore, individual differences in the cue effect on pain rating predicted the magnitude of participants' cue effect on the NPS response ($r(32) = 0.47$, $P = 0.007$; Fig. 2d). The effect of cue type was larger for more intense heat, as reflected in a Cue type × Temperature interaction ($t(33) = 2.0$, bootstrap $P = 0.04$, $d = 0.37$, CI = 0.25 to 5.1). Follow-up t-tests showed that NPS responses were significantly larger on high-pain than low-pain-cue trials for 49 °C heat ($t(33) = 3.1$, $P = 0.004$, $d = 0.53$), but not necessarily for 48° heat ($t(33) = 1.9$, $P = 0.07$, $d = 0.33$). Comparisons with neutral-cue trials showed that for 49 °C heat, there was a trend toward lower NPS responses following low versus neutral cues ($t(33) = 1.9$, $P = 0.06$, $d = 0.33$), but no significant difference between neutral and high cues ($t(33) = 1.4$, $P = 0.17$). For 48 °C heat, the NPS response on neutral-cue trials did not differ significantly from the NPS response on either low-pain-cue ($t(33) = 1.7$, $P = 0.11$) or high-pain-cue trials ($t(33) = 0.7$, $P = 0.49$).

Although the cue effect on the NPS response was numerically strongest in the first half of the test phase (Fig. 2e), there were no significant interactions between cue type and the linear ($t(33) = 1.5$, bootstrap $P = 0.12$, $d = 0.21$, CI = −0.13 to 0.03) or quadratic ($t(33) = 0.8$, bootstrap $P = 0.37$, $d = 0.15$, CI = −0.002 to 0.005) effects of time. Thus, we cannot say definitively whether the cue effect on the NPS extinguished over time, but it appears to persist throughout the test phase. Together, these findings show that the cues robustly and persistently modulated both behavioural and brain markers of pain.

Reciprocal positive associations between expectations and pain. The resistance to extinction of the cue effects on pain may be supported by reciprocal positive influences of expectations and pain perception on one another: expectations modify pain perception, and this modified perception drives subsequent expectations, resulting in the maintenance of expectations even when these do not reflect sensory input (Fig. 3a). To examine evidence for both halves of this putative reciprocal circuit, we first analysed the effects of expectations on subsequent pain ratings and NPS responses (left arrow in Fig. 3a), and then analysed the effects of pain ratings and NPS responses on subsequent cue-based expectations (right arrow in Fig. 3a). We controlled for cue type and temperature in all four analyses.

Consistent with the first half of the reciprocal circuit—expectation effects on subsequent pain—higher pain expectations predicted higher pain ratings in both studies ($t(27) = 8.8$, bootstrap $P < 0.001$, $d = 1.5$, CI = 0.37 to 0.61 and $t(32) = 7.4$, bootstrap $P < 0.001$, $d = 1.4$, CI = 0.25 to 0.41, in Study 1 and 2, respectively; Fig. 3b). Higher pain expectations also predicted higher NPS responses in Study 2 ($t(32) = 4.2$, bootstrap $P < 0.001$, $d = 0.71$, CI = 0.26 to 0.74; Fig. 3c). Furthermore, a multilevel-mediation analysis showed that trial-to-trial variation in NPS response formally mediated the effect of expectations on pain ratings (controlled for cue type and temperature; $P < 0.001$, $d = 0.70$). When controlled for NPS response, the effect of expectations on pain ratings remained highly significant (path c' (direct effect of expected pain on pain rating), $P < 0.001$, $d = 1.1$), implying a partial mediation.

Consistent with the second half of the reciprocal circuit—pain effects on subsequent expectations—higher pain ratings on a trial predicted higher pain expectations the next time the same cue was

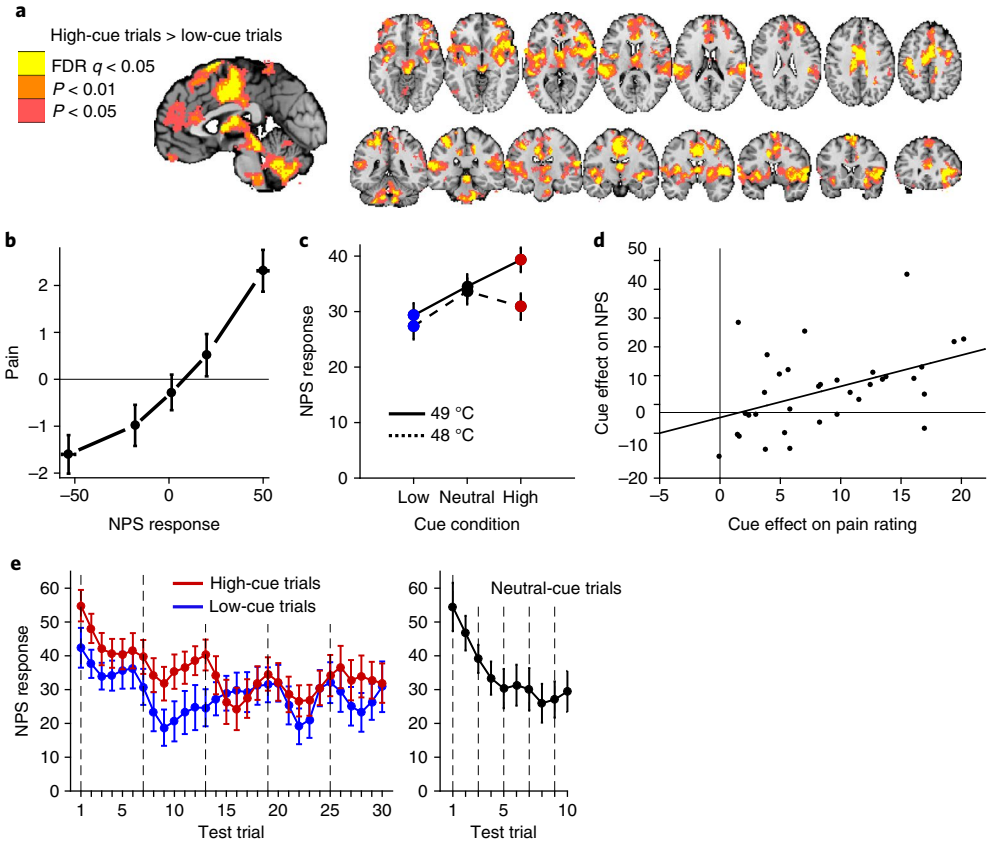

Fig. 2 | Cue effects on heat-evoked brain activity. a, Heat-evoked brain activity on high-pain-cue versus low-pain-cue trials. Coloured regions indicate stronger activity when heat was preceded by a high-pain than a low-pain cue. All coloured regions were significant at $q < 0.05$, false discovery rate (FDR)-corrected. For the purpose of display, we pruned the results using two additional, less conservative levels of voxel-wise threshold. **b**, Pain rating as a function of NPS response. We sorted each participant's pain ratings into five bins according to their single-trial NPS response (both mean-centred), and plotted the group-average data for each bin. Error bars indicate within-subject standard errors. **c**, NPS response as a function of stimulus temperature and cue type. **d**, Across-subject correlation (r) between the effects of cue type on pain rating and on the NPS response ($r(32) = 0.46$, $P = 0.007$, coefficient of determination (R^2) = 0.21, CI = 0.15 to 0.95). **e**, Average NPS response as a function of cue type and trial. Before plotting, the effect of temperature was regressed out and single-trial NPS responses were smoothed using locally weighted scatterplot smoothing. Vertical lines indicate the first trial of each scan run (heat was applied to a new skin site in each run). We note that the ten neutral-cue trials were evenly distributed amidst the low- and high-pain-cue trials (one neutral-cue trial was presented during each series of 3 low-pain and 3 high-pain cue trials). Error bars indicate between-subject standard errors. All plots are based on data from 34 participants.

presented ($t(27) = 8.0$, bootstrap $P < 0.001$, $d = 1.4$, CI = 0.17 to 0.29 and $t(32) = 4.5$, bootstrap $P < 0.001$, $d = 0.90$, CI = 0.10 to 0.22, in Study 1 and 2, respectively; Fig. 3d). In Study 2, there was a non-significant trend in the same direction for the NPS response: higher NPS responses on a trial predicted higher pain expectations the next time the same cue was presented ($t(32) = 1.2$, bootstrap $P = 0.10$, $d = 0.27$, CI = −0.001 to 0.01; Fig. 3e). Together, these results suggest that expectations modified pain perception, and that participants updated their expectations based on new pain experiences.

Confirmation bias in expectation updating. Bidirectional interactions between expectations and experience can contribute to persistent expectations in spite of inconsistent evidence. Importantly, however, the component of the pain response that is affected by prior expectations may differ from the component that drives expectation updating. Specifically, if expectancy-based pain modulation occurs at a processing stage subsequent to prediction-error signaling, expectancy effects on pain may not carry over to affect subsequent expectations (also see 'Discussion'). Moreover, even if there is a closed feedback loop between pain and expectations, this is sufficient for the creation of self-reinforcing expectancy effects only when experience fully conforms to the expectation. If experience adjusts only partially towards expectations, as was the case in our studies (Fig. 1d), we would expect a weakening of expectations and eventual extinction over time. Thus, to explain our findings of persistent cue effects on expected- and experienced-pain ratings, an additional mechanism is needed. One possible additional mechanism is a confirmation bias in learning[37]: stronger expectation updating when new experiences are consistent with the initial expectation of high or low pain than when they are inconsistent.

In our paradigm, a confirmation bias would be reflected in the strongest expectation updating when high- and low-pain cues are followed by, respectively, higher- and lower-than-expected pain. For conciseness, we refer to these as aversive and appetitive pain prediction errors, respectively, although this is not meant to suggest that lower-than-expected pain is processed by the same system as reward or pleasure. In reinforcement-learning models, the degree of expectation updating following prediction errors is controlled by the learning rate. Thus, a confirmation bias can be statistically formalized as an interaction between cue type (high- versus low-pain cue) and prediction error sign (aversive versus appetitive) on learning rate. Prediction errors and learning rate are typically inferred

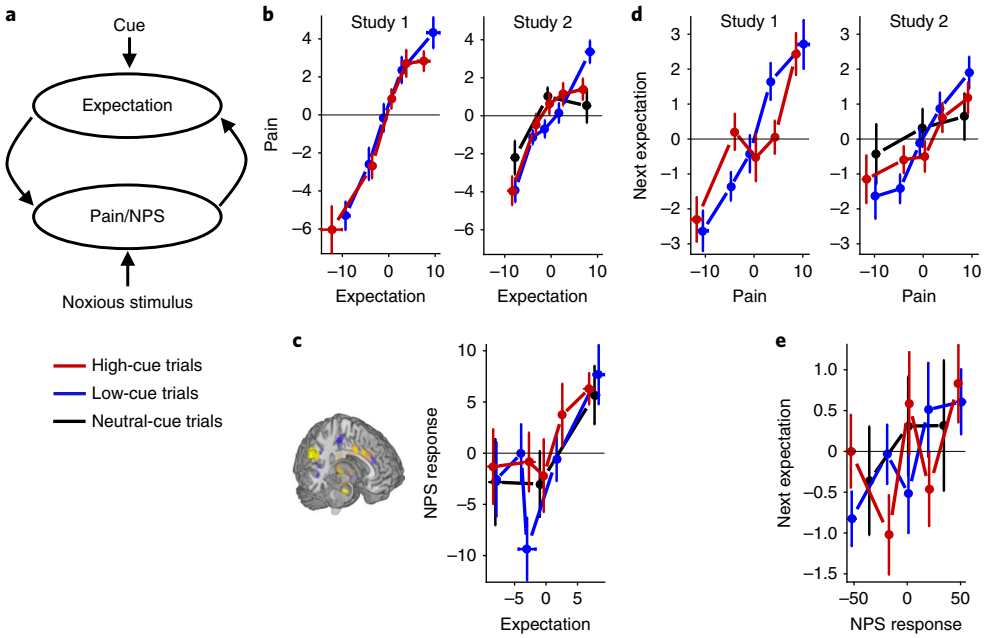

Fig. 3 | Bidirectional effects of expectations and pain on one another. a, Dynamic feedback circuit through which expectations can become self-reinforcing. **b**, Pain rating as a function of expected pain and cue type. We sorted each participant's pain ratings for the low-pain-cue and high-pain-cue trials into five bins, according to their trial-specific expected-pain ratings (both mean-centred), and plotted the group-average data for each bin. Because there were fewer neutral-cue trials, we used three bins for the neutral-cue condition. We note that we binned the data for plotting purposes, but used single-trial measures in our statistical analyses. **c**, NPS response as a function of expected pain and cue type. **d**, Cue-based pain expectation as a function of cue type and the previous pain rating for that cue. **e**, Cue-based pain expectation as a function of cue type and the previous NPS response for that cue. All error bars indicate between-subject standard errors. All plots for Study 1 and 2 are based on data from 28 and 33 participants, respectively.

from behavioural or autonomic data, but the trial-specific expectation and pain ratings in our paradigm provide direct estimates of these variables on each trial, requiring fewer assumptions about latent variables and providing an empirical estimate of learning in each condition. Specifically, we defined an estimated prediction error ($\hat{\delta}$) on each trial t as the difference between the ratings of pain (P) and expected pain (E): $\hat{\delta}_t = P_t - E_t$. Additionally, we defined an estimate of trial-specific learning rate ($\hat{\alpha}$) as the change in expected-pain rating across two successive presentations of the same cue c (trials t and t'), divided by the most recent prediction error for that cue: $\hat{\alpha}_{c,t} = (E_{c,t'} - E_{c,t}) / \hat{\delta}_t$. We report additional model-based analyses in the 'Computational models' section.

In Study 1, estimated learning rate did not differ between high- and low-pain cues ($t(25) = 0.2$, bootstrap $P = 0.81$, $d = 0.05$, CI = −0.12 to 0.16), but was marginally higher for aversive than appetitive prediction errors ($t(25) = 1.2$, bootstrap $P = 0.07$, $d = 0.31$, CI = −0.03 to 0.33). Importantly, learning rates were highest for appetitive prediction errors following low-pain cues and for aversive prediction errors following high-pain cues, resulting in an interaction between cue type and prediction error sign ($t(25) = 2.5$, bootstrap $P < 0.001$, $d = 0.68$, CI = 0.08 to 0.28; Fig. 4a), consistent with a confirmation bias. Study 2 confirmed these findings. In Study 2, participants showed higher estimated learning rates for high- than low-pain cues ($t(20) = 2.5$, bootstrap $P < 0.001$, $d = 0.49$, CI = 0.17 to 0.95) and for aversive than appetitive prediction errors ($t(20) = 2.1$, bootstrap $P = 0.005$, $d = 0.40$, CI = 0.08 to 0.94). As in Study 1, learning rates were highest for appetitive prediction errors following low-pain cues and aversive prediction errors following high-pain cues, resulting in an interaction between cue type and prediction error sign ($t(20) = 2.4$, bootstrap $P < 0.001$, $d = 0.46$, CI = 0.16 to 1.1; Fig. 4a). Thus, learning rates were consistent with a confirmation bias in both studies. We also performed the same analysis on raw expectancy updates (not divided by prediction error), which yielded similar results (Supplementary Results and Supplementary Figure 1), demonstrating that the results are not driven disproportionately by high learning rates on trials with very small prediction errors.

The previous analysis considered the sign, but not the magnitude, of prediction errors. Also, a limitation of the previous analysis is that there were relatively few high-pain-cue trials with aversive prediction errors (on average 9.1 and 4.2 trials per participant in Study 1 and 2, respectively) and low-pain-cue trials with appetitive prediction errors (on average 8.7 trials per participant in both studies). Two participants in Study 1 and 12 participants in Study 2 never experienced aversive prediction errors on high-pain-cue trials, and hence were excluded from the previous analysis (leaving 26 and 21 participants in Study 1 and 2, respectively).

To address these issues, we conducted an additional analysis in which we tested for effects of cue type and signed prediction error magnitude on (signed) expectation updating. All participants were included in this analysis. As expected, expectation updating increased with increasing prediction error magnitude ($t(27) = 12.7$, bootstrap $P < 0.001$, $d = 2.3$, CI = 0.30 to 0.42 and $t(32) = 7.1$, bootstrap $P < 0.001$, $d = 1.6$, CI = 0.22 to 0.34 in Study 1 and 2, respectively; Fig. 4b). In addition, there was a main effect of cue type ($t(27) = 4.8$, bootstrap $P < 0.001$, $d = 1.6$, CI = 0.76 to 1.2 and $t(32) = 5.0$, bootstrap $P < 0.001$, $d = 1.1$, CI = 0.86 to 1.6 in Study 1 and 2, respectively), indicating that high-pain cues, relative to low-pain cues, caused an upward shift in expectation updating. There was no Prediction error × Cue type interaction (bootstrap $P > 0.17$ in both studies), suggesting that this cue effect on expectation updating was independent of prediction error magnitude. Together, these effects reflect that upward updating of pain expectations following aversive prediction errors was stronger for high- than low-pain cues, while downward updating of pain expectations following appetitive prediction errors was stronger for low- than high-pain cues (Fig. 4b). Thus, the cues

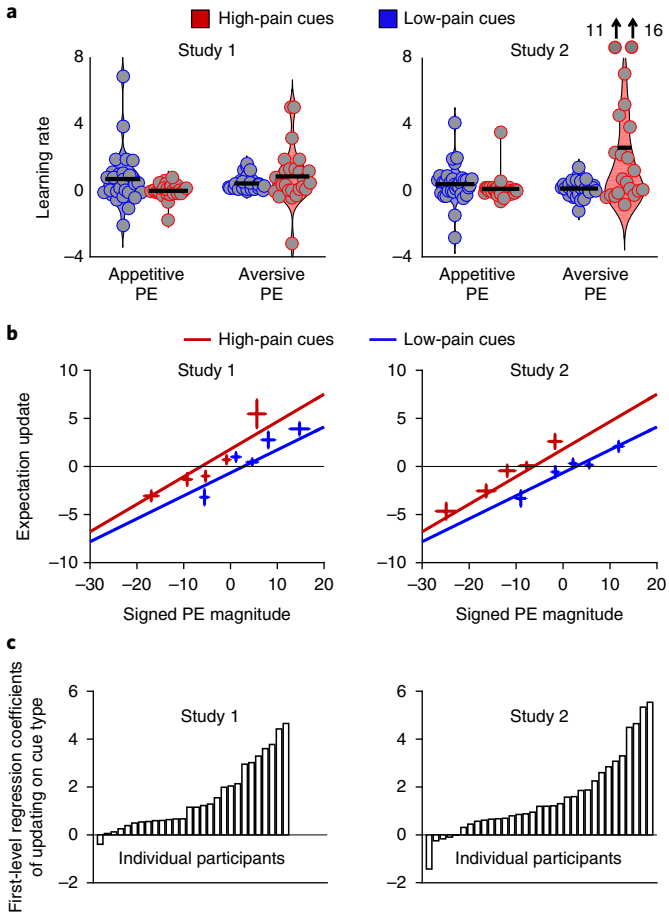

Fig. 4 | Confirmation bias in expectation updating. a, Estimated learning rate as a function of prediction error sign and cue type in each study. PE, prediction error. The grey dots indicate individual participants, and the horizontal black lines are the mean values in each condition. The two individual data points at the top of the last condition (learning rates for aversive prediction errors on high-pain-cue trials) in Study 2 fall outside the range of the figure; estimated learning rates for these points are 11 and 16. For Study 1, the first three conditions include data from 28 participants, and the last condition includes data from 26 participants. For Study 2, the first three plots include data from 33 participants, and the last plot includes data from 21 participants. Fewer participants contributed to the last condition because some participants never experienced aversive prediction errors on high-pain-cue trials. **b**, Expectation updating as a function of signed prediction error magnitude and cue type in each study. Negative and positive prediction errors indicate lower- and higher-than-expected pain, respectively. Negative and positive expectation updates indicate decreases and increases in pain expectations, respectively. We sorted each participant's low- and high-pain-cue trials into five bins according to signed prediction error magnitude, and plotted the group-mean signed expectation updates for each bin. Lines show linear fits to unbinned single-trial data. We note that there was a significant main effect of cue type but no significant interaction between prediction error and cue type. Plots for Study 1 and 2 are based on data from 28 and 33 participants, respectively. Error bars indicate between-subject standard errors. **c**, Effect of cue type on expectation updating in each participant (first-level regression coefficients).

influenced expectation updating in accordance with a confirmation bias. Finally, the effect of the cues on expectation updating varied substantially across participants (Fig. 4c).

Computational models. To formalize and quantify the latent processes that we hypothesized to underlie the observed cue effects, we developed two computational models: a reinforcement learning model and a Bayesian model. The two models embody the same ideas from different perspectives and are formally closely related; hence they can be seen as complementary rather than competing. The reinforcement learning model explains pain perception as a mechanistic process involving prediction and error correction, whereas the Bayesian model explains it at a rational level in terms of probabilistic inference. Full descriptions of the models, and their equations, can be found in the Methods; here we provide a brief overview of their main points. The models observe the sequence of cues and noxious stimuli presented to participants; each cue is initially associated with a low or high pain value. Both models contain (1) a perceptual inference mechanism, which models the pain experienced on a given trial as a weighted average of the sensory (noxious) input and cue-based pain expectation, and (2) a learning mechanism, which governs how each new pain experience is used to update the predictive value (expectation) of the cue that preceded it (Fig. 5). Model 1, the reinforcement learning model, assumes that participants represent expectations and pain outcomes as point values that are updated based on prediction error, as in standard reinforcement learning models. It has three free parameters (which are estimated for each participant) that control the relative weighting of expectations and sensory input during perceptual inference, and the learning rates for cue–pain associations following cue-consistent and cue-inconsistent prediction errors (see below). Model 2, the Bayesian model, assumes that participants track not only the expected values of cue–pain associations but also their uncertainties (variances), and it determines the influence of expectations

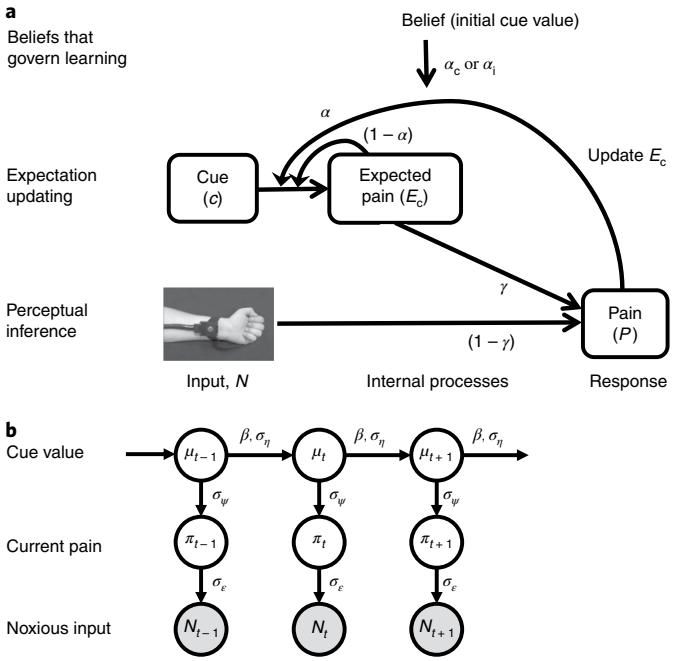

Fig. 5 | Computational models capturing effects of cue-based expectations on pain and confirmation bias on expectation updating. a, Reinforcement learning model (Model 1). Perceptual inference within a trial combines expectations with noxious input to determine perceived pain. The γ parameter controls the relative impact of these two sources. Learning between trials involves updating the expectation for the current cue toward the current perceived pain. Experience-resistant expectations are modelled by assuming different learning rates, the α_c and α_i parameters, when the direction of prediction error is, respectively, consistent and inconsistent with the cue's initial low or high pain association. **b**, Bayesian model (Model 2). Pain perception and expectation are products of Bayesian inference with respect to a generative model of the task environment that extends the classic Kalman filter. We note that arrows in this diagram indicate statistical dependencies in the subject's generative model, not dynamics of the subject's state of knowledge as in **a**. Under the generative model, the mean threat level signalled by each cue (μ_c, index c suppressed in figure) drifts randomly from trial to trial, with step size determined by the σ_η parameter. The current threat level (or objectively correct pain level, π_t) on any trial t deviates randomly from μ_t, with standard deviation equal to the σ_ψ parameter. The noxious input (N_t) is a noisy indicator of π_t, with standard deviation equal to the σ_ε parameter. Inference within a trial (not shown) combines the current belief about μ_t with the observed value of N_t to estimate the current value of π_t; this estimate is the subject's experienced level of pain (P_t). Inference across trials combines the beliefs about μ_t and π_t to estimate μ_{t+1}; this is the subject's reported expectation the next time this cue is presented (E_{t+1}). Experience-resistant expectations are modelled by letting cue–pain associations drift in the directions of their initial values between trials, to a degree governed by the β parameter. Model 2 is formally nearly equivalent to Model 1, except that it assumes γ and α adapt from trial to trial to reflect the subject's current level of uncertainty (equations (23), (25) and (28)).

on pain and the learning update following a given trial based on these uncertainty estimates. Both expected values and variances are updated over time.

Model 1: Reinforcement learning model. Model 1 is a mechanistic model expressed in the framework of reinforcement learning. The perceptual inference component of this model computes the pain evoked by a noxious stimulus as a weighted average of the current sensory input (stimulus intensity) and the current cue-based expectation. The weighting of sensory input versus expectation is controlled by a parameter $\gamma \in [0,1]$, which is constant across trials. Higher γ yields a stronger impact of expectations on pain. The learning component of this model assumes that each new pain experience triggers the updating of the relevant cue's pain expectation in proportion to the prediction error (pain minus expected pain), according to a standard reinforcement-learning algorithm (delta rule)[32]. A learning-rate parameter, α, controls the degree of expectation updating: high values of α result in strong updating toward the latest pain experience, whereas low values of α result in more slowly varying expectations. To model confirmation bias, we assume that on each trial one of two different learning rates is used—α_c or α_i—depending on whether the sign of the prediction error is consistent or inconsistent, respectively, with the cue's initial low or high pain association (as acquired in the learning phase). If α_c is higher than α_i this implies a confirmation bias. Thus, this model has three free parameters: γ, α_c and α_i. Parameter γ controls the impact of the current expectation on pain, and α_c and α_i govern learning and its dependence on initial beliefs.

Model 2: Bayesian model. Model 2 considers pain perception and learning in terms of Bayesian inference, representing expectations and perceived pain as probabilistic beliefs obeying Gaussian probability distributions[16,54–56]. Under this model, perceived pain is treated as the subject's inferred belief regarding how threatening (in terms of potential tissue damage) a situation truly is. The perceptual inference component of the model computes this belief on each trial in a Bayesian fashion, by combining prior expectations based on environmental cues with current nociceptive input. The mean of the resulting posterior belief is a weighted average of these two sources, as in Model 1, but their relative impact depends on their precisions (inverse variances), such that more precise expectations are weighted more heavily. The learning component of the model is based on a Kalman filter, which tracks estimates of the mean level of potential harm signaled by each cue (that is, the objectively correct cue–pain association) as well as the precision of those estimates[57,58]. The effective learning rate on each trial depends on the precision of

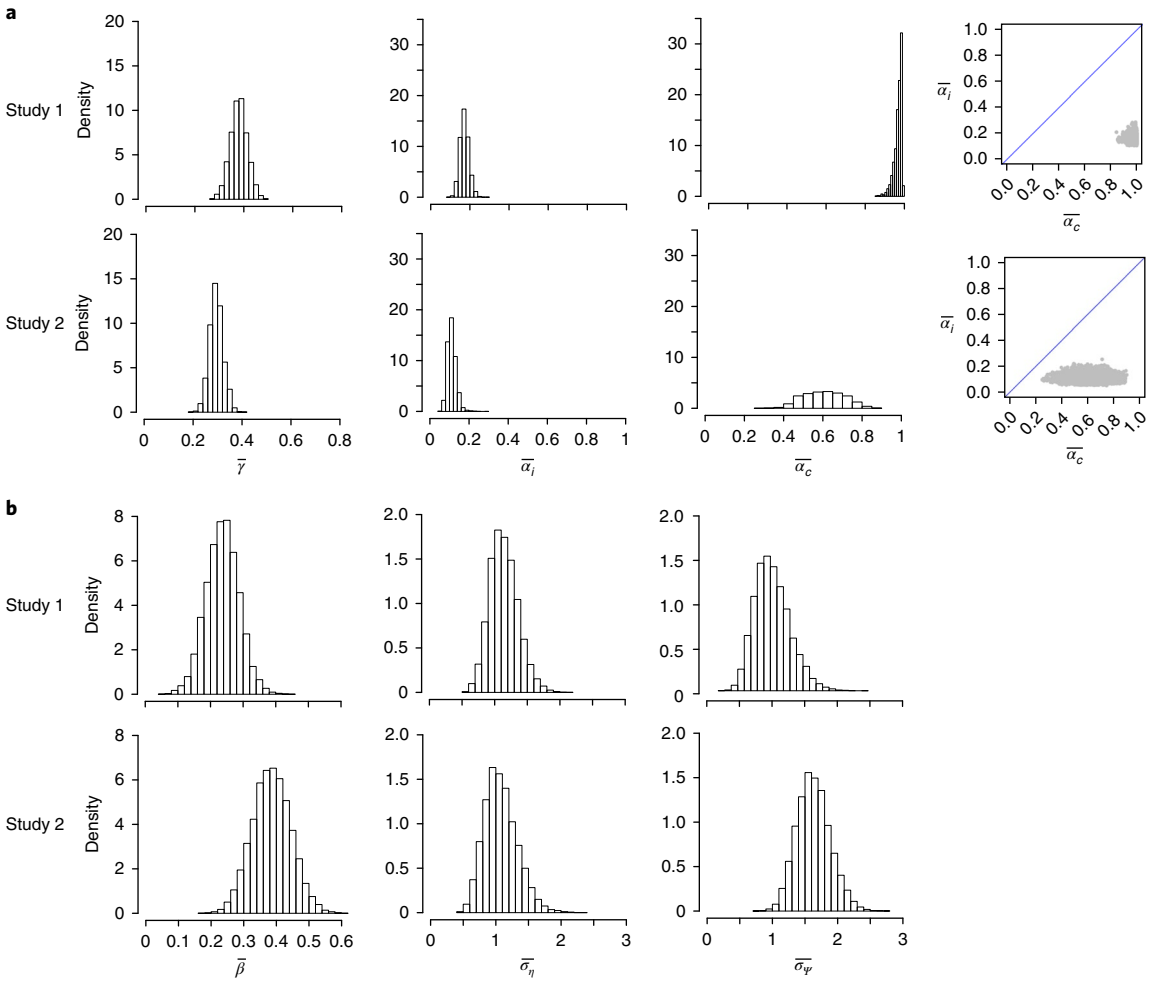

Fig. 6 | Posterior distributions for the group-level means of the models' parameters. a, Model 1: γ controls the impact of expectations versus nociceptive input on pain (higher values cause a stronger weighting of expectations); α_c and α_i are learning rates for cue-consistent and cue-inconsistent prediction errors. The rightmost panels are joint density plots of $\tilde{\alpha}_c$ and $\tilde{\alpha}_i$ (dots are samples from the MCMC), showing that $\tilde{\alpha}_c$ is reliably greater than $\tilde{\alpha}_i$. **b**, Model 2: β controls the drift of expectations towards (if $\beta > 0$) or away from (if $\beta < 0$) their initial values after each update; σ^2_ψ is the assumed variance in pain on any given trial; σ^2_η is the assumed variance of the random walk process (random variation in pain across trials). σ^2_ψ and σ^2_η are estimated relative to the variance of contributions to noxious input unrelated to pain (σ^2_ε, which was fixed at 1 to eliminate redundancy in model parameters).

the expectation at the onset of that trial, with less precise expectations leading to greater updating based on current perceived pain. In addition, the Kalman filter assumes that the objective cue–pain associations vary over time according to a Gaussian random walk process, adding some uncertainty to the cue-based pain expectations after each trial. To model a persistent influence of initial expectations, we incorporated an assumption that the dynamics of the random walk can be biased toward the values of the initial cue–pain associations (see 'Comparison of Models 1 and 2' in the Methods for an explanation of how this relates to biased learning). The amount of bias is controlled by free parameter $\beta \in [-1, 1]$: If β equals 0 the random walk process has no directional bias, values of β above 0 yield a drift toward the initial cue-based pain expectations (bias toward initial beliefs), and values of β below 0 yield a drift in the opposite direction of the initial expectations (bias away from initial beliefs). Besides β, this model has three parameters characterizing the subject's belief about the generating process: the variance of the random walk process (σ^2_η), the variance in level of harm on a given trial around the average predicted by the current cue (σ^2_ψ), and the variance of noise in the noxious input (σ^2_ε). Only the ratios among these three parameters matter; hence we fixed σ^2_ε to 1 and estimated σ^2_η and σ^2_ψ as free parameters (see Methods). Thus, like Model 1, Model 2 also has three free parameters.

Parameter estimates. To obtain quantitative estimates of the model parameters, we fitted each model (and several reduced versions; see below) to participants' trial-to-trial sequences of pain and expected-pain ratings, using hierarchical Bayesian parameter estimation (see Methods). The hierarchical procedure assumes that every participant has a different set of model parameters, drawn from some population distribution. As we are primarily interested in average participant behaviour, the primary variables of interest are the means of the population distributions, which we denote with overbars (for example, $\bar{\gamma}$ denotes the population mean for γ in Model 1). Posterior distributions of the population means for all model parameters are shown in Fig. 6. The estimated population distributions of each parameter are shown in Supplementary Fig. 2.

The posterior distribution of $\bar{\gamma}$ in Model 1 (Study 1: median = 0.38, 95% credible interval (CR) = 0.32 to 0.45; Study 2: median = 0.29, 95% CR = 0.24 to 0.35) suggests that perceived pain is jointly determined by expectations and noxious input, and that, for the average participant, noxious input is weighted roughly twice as strongly

Table 1 | $\log_e$[Bayes factor] for all model comparisons

Study 1

	M1	M1a	M1b	M1c	M2	M2a	M2b	M2c
M1	0	391	129	529	57	459	197	592
M1a		0	−262	138	−334	68	−194	201
M1b			0	400	−72	330	69	463
M1c				0	−472	−70	−331	63
M2					0	402	141	535
M2a						0	−261	133
M2b							0	395
M2c								0

Study 2

	M1	M1a	M1b	M1c	M2	M2a	M2b	M2c
M1	0	263	86	334	36	291	153	399
M1a		0	−177	71	−227	28	−109	137
M1b			0	248	−50	205	68	313
M1c				0	−298	−43	−180	65
M2					0	255	118	363
M2a						0	−137	108
M2b							0	246
M2c								0

We log-transformed the Bayes factors such that a value of 0 indicates that the data are equally likely to occur under both models. Positive values indicate evidence in favour of the model in the row over the model in the column. M1 and M2 are Model 1 and 2, respectively, and a, b and c refer to the reduced models that do not include (a) expectancy-based pain modulation, (b) a confirmation bias and (c) either of those components.

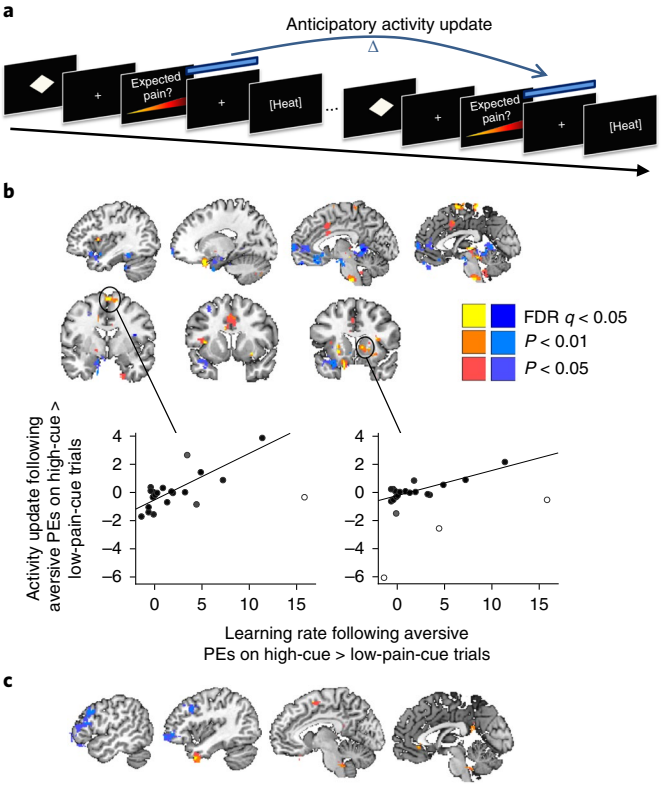

Fig. 7 | Confirmation bias in the updating of pain-anticipatory brain activity. a, We computed the change in anticipatory activity across successive trials in which the same cue was presented, separately for trials with higher- and lower-than-expected pain (aversive and appetitive prediction errors, respectively), and tested for effects of cue type. **b**, Individual differences in confirmation bias on estimated learning rate predict participants' confirmation bias on anticipatory activity updating following aversive prediction errors. Yellow/red colours indicate positive across-subject correlations between increases in anticipatory activity following aversive prediction errors on high- versus low-pain-cue trials and estimated learning rate following aversive prediction errors on high- versus low-pain-cue trials. Blue colours indicate negative correlations. Cluster statistics can be found in Supplementary Table 2. Scatterplots illustrate the correlations in sensorimotor cortex and right striatum. The shading of the points is proportional to each observation's weight in robust regression; white dots indicate outliers down-weighted by the algorithm. **c**, Individual differences in confirmation bias on estimated learning rate predict participants' confirmation bias on anticipatory activity updating following appetitive prediction errors. Yellow/red colours indicate positive across-subject correlations between increases in anticipatory activity following appetitive prediction errors on low- versus high-pain-cue trials and estimated learning rate following appetitive prediction errors on low- versus high-pain-cue trials. Blue colours indicate negative correlations. All coloured regions were significant at $q < 0.05$, FDR-corrected. For display purposes, we show the extent of results surrounding FDR-corrected peaks at $P < 0.01$ and $P < 0.05$ uncorrected. Cluster statistics can be found in Supplementary Table 3. Panels **b** and **c** are based on data from 21 participants.

as expectations. Furthermore, $\overline{\alpha}_c$ was considerably higher than $\overline{\alpha}_i$, consistent with a confirmation bias on learning rate: for Study 1, posterior medians were 0.98 for $\overline{\alpha}_c$ (95% CR = [0.93,1.0]) and .16 for $\overline{\alpha}_i$ (95% CR = [0.13,0.21]); for Study 2, posterior medians were 0.63 for $\overline{\alpha}_c$ (95% CR = [0.42,0.80]) and .11 for $\overline{\alpha}_i$ (95% CR = [0.07,0.16]). At the individual level, the median of the posterior distribution was higher for α_c than for α_i in 100% and 91% of the participants in Studies 1 and 2, respectively (sign tests, P values < 0.001).

In Model 2, the relative impact of expectations and noxious input on perceived pain (governed by γ in Model 1) is determined on a trial-to-trial basis by the current level of uncertainty about the objective cue-pain association, and by σ_ε^2 and σ_ψ^2 (see 'Comparison of Models 1 and 2' in the Methods). The average weighting of expectations versus noxious input in Model 2 (the grand-average estimate of γ_i, as defined in equation (23), using the medians of the group-level parameter estimates) was 0.40 and 0.32 in Study 1 and 2, respectively, resembling the median posterior estimates of $\overline{\gamma}$ in Model 1. Parameter β in Model 2 indexes the degree to which expectations drift towards (if $\beta > 0$) or away from (if $\beta < 0$) their initial values after updating, which is a Bayesian way to capture biased learning. In both studies, 100% of the numerical estimate for the posterior distribution of $\overline{\beta}$ lay above 0 (Study 1: median = 0.24, 95% CR = 0.17 to 0.32; Study 2: median = 0.36, 95% CR = 0.27 to 0.46), consistent with a persistent contribution of initial beliefs.

Model comparison. To further examine evidence for expectation-based pain modulation and a confirmation bias in learning, we fitted three reduced versions of each model that omitted (a) expectancy-based pain modulation ($\gamma = 0$ in Model 1a, and $\sigma_\varepsilon^2 = 0$ in Model 2a), (b) biased learning ($\alpha_c = \alpha_i$ in Model 1b, and $\beta = 0$ in Model 2b), or (c) either of those processes (Model 1c and Model 2c). We note that when σ_ε^2 is 0 (Models 2a and 2c), there is no unexplained error in the noxious input (equation (7)) and therefore no possibility of expectations affecting the inferred pain (equation (13)). When σ_ε^2 is 0, only the ratio between σ_ψ^2 and σ_η^2 matters. Therefore we fixed σ_ψ^2 to 1 and estimated σ_η^2 as a free parameter for Models 2a and 2c. We compared the different models using Bayes factors, which quantify how much more likely the data are to occur under one model

compared to another model (Table 1). For both studies, Bayes factors clearly favoured the full versions of Models 1 and 2 over all their reduced versions, indicating that the data are best explained by models including both expectation-based pain modulation and a confirmation bias in learning. In addition, models with only confirmation bias and only expectancy-based pain modulation were both favoured over reduced models with neither component. Also note that Model 1 (the reinforcement learning model) was favoured over Model 2 (the Bayesian model), suggesting that participants did not adapt their learning rates or pain inferences over time as a function of the precision of their expectations. Of course, this result bears only on the utility of these particular models, and does not imply that reinforcement learning models fit better than Bayesian models in general.

Model simulations. To assess whether our models capture the key features of the behavioural results, we simulated expectation and pain ratings according to our full and reduced models, using the best-fitting parameter values (Supplementary Fig. 3). Only models including both expectancy-based pain modulation ($\gamma > 0$ and $\sigma_\varepsilon^2 > 0$ in Models 1 and 2, respectively) and biased learning ($\alpha_c > \alpha_i$ and $\beta > 0$ in Models 1 and 2, respectively) capture the key features of the data: a persistent cue effect on both expectation and pain ratings, with a stronger cue effect on expectations than on pain. Although expectations become less extreme over time, indicating that learning is taking place, expectations and pain never fully converge. In contrast, models without expectancy-based pain modulation predict that the cues do not affect pain, and models without biased learning predict that expectations and pain converge to the actual noxious-input intensity. Thus, none of these reduced models fits the pattern of data we observed in these studies, qualitatively or quantitatively, indicating that both expectancy-based pain modulation and biased learning towards initial expectations occur.

Possible persistent effects of initial beliefs on pain. The parameter estimates and model comparisons support the existence of a confirmation bias in learning. However, a potential alternative explanation for these results is a persistent influence of participants' initial expectations on pain perception: the assimilation of pain towards the initial expectation on all trials would reduce cue-inconsistent prediction errors, which could result in an apparent confirmation bias without affecting learning rate itself.

To test this alternative explanation, we extended the perceptual-inference component of Model 1 by adding an effect of the initial expectation acquired during the learning phase, such that pain was computed as a weighted average of the initial expectation, the current (updated) expectation and the noxious input. The weights of these three terms were controlled by parameters λ, γ and $(1-\lambda-\gamma)$, respectively, where γ, $\lambda \geq 0$ and $(\gamma+\lambda) \in [0,1]$. We refer to this extended model as Model 1^+. We also defined Model $1b^+$ as the corresponding extension of Model 1b. Thus Models 1^+ and $1b^+$ both include a persistent effect of initial beliefs on pain, and they differ in that Model 1^+ includes biased learning ($\alpha_c > \alpha_i$) whereas Model $1b^+$ does not ($\alpha_c = \alpha_i$). Comparing these models provides a test of whether there was still evidence for a confirmation bias in learning when allowing for a persistent effect of initial beliefs on pain. Bayes factors strongly favoured Model 1^+ over Model $1b^+$ in both studies ($\log_e$[Bayes factor] = 78 and 102 for Study 1 and 2, respectively). In addition, Model 1^+ was favoured over our original Model 1 in Study 2 ($\log_e$[Bayes factor] = 8) but not in Study 1 ($\log_e$[Bayes factor] = −12). Supplementary Table 1 reports Bayes factors for comparisons of all versions of Model 1. Thus, these results provide mixed support for a persistent effect of initial beliefs on pain perception. More importantly, they strengthen our conclusions regarding confirmation bias in learning rate, showing that this mechanism is strongly supported regardless of any direct, persistent effects of initial beliefs on pain.

Confirmation bias in the updating of pain-anticipatory brain activity. When pain expectations are updated during learning, this probably causes changes in pain-anticipatory brain responses. Specifically, cues followed by more pain than expected (aversive prediction errors) should increase pain-anticipatory activity on subsequent presentations. Conversely, cues followed by less pain than expected (appetitive prediction errors) should reduce pain-anticipatory activity on subsequent presentations. If confirmation biases are present, this updating of anticipatory activity should be particularly strong when the sign of prediction errors is consistent with the cue's original low or high pain association. Therefore, we hypothesized that the confirmation bias in expectation updating would be paralleled by a confirmation bias in the updating of cue-specific pain-anticipatory brain activity (during the interval between expectation rating and heat onset).

Within-person parametric modulation analyses revealed that anticipatory brain activity increased in proportion to pain expectations in several regions, including the supplementary motor area, sensorimotor cortex, anterior midcingulate cortex, anterior insula and frontal operculum, temporal pole, and inferior cerebellar vermis (Supplementary Fig. 4; $P < 0.001$, uncorrected). While anticipatory activity in some regions (anterior midcingulate cortex, anterior insula and frontal operculum, temporal pole) is predicted from prior studies[6,59–62], the activation did not reach significance in whole-brain false discovery rate (FDR) correction. Preliminary evidence for activation of these regions may reflect increased anticipatory anxiety and alertness with increasing pain expectations.

To test for a confirmation bias in the updating of cue-specific pain-anticipatory brain activity, we estimated an fMRI activation map for each pain-anticipation period using single-trial modelling[49,63,64]. We then computed the change in anticipatory activity across successive presentations of the same cue, resulting in anticipatory activity 'update' images (Fig. 7a). To examine evidence for a confirmation bias in the updating of anticipatory activity, we computed contrast images for (1) activity updates following aversive prediction errors that were stronger following high-pain cues than following low-pain cues; and (2) activity updates following appetitive prediction errors that were stronger following low-pain cues than following high-pain cues. To test whether cue effects on activity updates were related to individual differences in susceptibility to confirmation bias, we added a regressor coding for each participant's confirmation bias on learning rate (computed directly from the rating data; centred) for the corresponding contrast. Participants who never experienced aversive prediction errors on high-pain-cue trials were excluded from this analysis (leaving 21 participants).

At the group-mean level, the cues did not significantly affect anticipatory activity updates following aversive prediction errors after whole-brain FDR correction. However, individual differences in the confirmation bias on learning rate predicted activity for this contrast in several brain regions ($q < 0.05$, FDR-corrected). In participants with larger confirmation biases, aversive prediction errors following high- relative to low-pain cues resulted in greater increases in anticipatory activity in the anterior midcingulate cortex, sensorimotor cortex, mid-insula, striatum (caudate and putamen, but not the nucleus accumbens), a midbrain region encompassing the periaqueductal gray, a region in the pons, and the inferior cerebellar vermis (Fig. 7b; Supplementary Table 2). Note that several, but not all, of those regions overlapped with regions tracking pain expectations (Supplementary Fig. 4). Negative correlations (shown in blue in Fig. 7b) were found in the ventromedial prefrontal cortex (PFC), as well as the dorsal anterior insula and the superior cerebellum. Thus, in individuals with a stronger confirmation bias, aversive prediction errors following high-pain cues (relative to low-pain cues) resulted in increased subsequent anticipatory responses in regions broadly associated with threat and anxiety (anterior midcingulate cortex, insula, periaqueductal gray), and decreased subsequent

anticipatory responses in regions commonly associated with appetitive value (ventromedial PFC; see Discussion).

In a complementary analysis, we tested for an effect of cue type on the updating of anticipatory activity following appetitive prediction errors. If learning from aversive and appetitive prediction errors is subserved by one underlying threat/safety system, this analysis should reveal opposite activity patterns to those reported in the previous section. Alternatively, if there are different systems for threat and safety learning, this analysis may reveal different regions. At the group-mean level, the cues did not significantly affect anticipatory activity updates following appetitive prediction errors (FDR-corrected). However, individual differences in the confirmation bias on learning rate predicted activity for this contrast in several brain regions. In participants with a greater confirmation bias, appetitive prediction errors following low-pain cues (relative to high-pain cues) resulted in a greater increase in anticipatory activity in the ventromedial PFC and posterior cingulate cortex—both often considered 'default mode' regions—as well as the orbitofrontal cortex, temporal pole and pre-supplementary motor area (Fig. 7c; Supplementary Table 3). Negative correlations (blue in Fig. 7c) were found in the lateral PFC. Thus, the ventromedial PFC showed parallel learning effects in analyses of both aversive and appetitive prediction errors. In individuals with a stronger confirmation bias, anticipatory ventromedial PFC responses decreased more following aversive prediction errors on high-pain-cue than low-pain-cue trials, and increased more following appetitive prediction errors on low-pain-cue than high-pain-cue trials. In other respects, however, effects of cue type on anticipatory-activity updating were largely distinct for aversive and appetitive prediction errors.

Discussion

Positive feedback loops between expectations and experiences can create self-fulfilling prophecies—persistent beliefs, behaviours, and decisions that are driven by initial expectations and are resistant to corrective experience. Although such reciprocal systems may operate in many aspects of human behaviour[20,65], empirical studies investigating them are rare, and their underlying mechanisms largely unknown.

Our findings provide evidence that cue-based expectations about pain can become self-reinforcing through their combined effects on perception and learning. First, we identified reciprocal effects of expectations and pain responses on one another: expectations influenced reported pain and pain-specific brain activity, and these pain responses in turn influenced subsequent expectations. Second, expectation updating was strongest when the direction of prediction errors was consistent with the initial high or low pain value of the preceding cue. Third, computational modelling results provided additional evidence that self-reinforcing expectancy effects on pain result from a combination of expectancy-based pain modulation and a confirmation bias in learning. Fourth, individual differences in confirmation bias in learning predicted confirmation bias in the updating of pain-anticipatory brain activity in regions associated with threat and affective value.

Our results contribute to a basis for understanding self-fulfilling predictions in various domains. In principle, either expectancy effects on perception or a confirmation bias in learning alone, if sufficiently strong, would be sufficient to maintain expectations for a period of time. If perception fully assimilates towards expectations, we always experience what we expect, and learning may be short-circuited. Similarly, if we do not at all learn from evidence that does not confirm what we expect, our initial expectations will be maintained indefinitely. However, extreme biases in perception and learning are unlikely to be functionally advantageous. Complete adjustment of perception to expectations is a form of hallucination, and would prevent the detection of unexpected harmful or rewarding events. Similarly, a complete suppression of learning from events that do not confirm expectations would prevent adaptive future responding to those events. Our results show that, when combined, even moderate degrees of perception modulation and biased learning can produce persistent expectations, with substantial impacts on perceived pain.

These results complement those of previous computational-modelling studies in other domains that explained effects of prior information on people's choice behaviour by a modulation of outcome evaluation or a confirmation bias in learning[36–38,41]. Although these two mechanisms are computationally distinct—affecting the input to the learning process and the learning rule itself, respectively—they make similar predictions about trial-to-trial dynamics of expected value, and hence are difficult to dissociate based on choice data alone. By directly measuring trial-to-trial dynamics of both expected and perceived outcomes, our paradigm enables the dissociation of these two effects.

Our results also contribute to the discussion about predictive coding and the strength, importance and locus of 'top-down' influences on perception. Previous studies have shown that higher-level cognitive processes, such as expectations and attention, can affect perceptual judgments of stimuli in visual[10–12], pain[23], auditory[66] and taste[13] domains. It has recently been questioned whether these findings—specifically those in the visual domain—really reflect the modulation of perception, or merely the interpretation of perceived events[67]. The present study contributes to this debate by showing that expectations modulate pain signalling in the central nervous system in functionally important, and clinically relevant, ways. The NPS is a cerebral pain signature that is strongly affected by noxious-stimulus intensity[49,53] and by drugs that are thought to act in part at a spinal level[53,68]. In contrast, changes in reported pain due to reappraisal[49], pain controllability[69,70], and reward manipulations[71] are not reflected in the NPS response, suggesting that the NPS is largely insensitive to higher-order cognitive processes. Thus, our finding that cue-based expectations modulate the NPS response implies that they affect the cerebral processes that give rise to pain.

With regard to 'predictive coding', recent work has proposed that pain perception results from the Bayesian integration of afferent nociceptive input with 'top-down' expectations and/or other informational cues[16,54–56], as is also assumed by our Bayesian model. According to these accounts, the precision of each information source determines its influence on perceived pain and the degree of expectation updating; in particular, very precise expectations strongly shape pain and reduce experience-driven expectation updating. Our data are compatible with this view, but they suggest that prior expectations have an additional effect not captured in standard Bayesian or non-Bayesian learning models. Specifically, our data suggest that initial beliefs about cue-pain associations modify learning rates, creating long-lasting confirmation biases that impede extinction.

Such effects have a history of study in social psychology, under the label of 'attribution', which relates to a person's theory about the causal structure underlying experienced associations[72,73]. In short, attributions govern what a person 'takes away' or learns from experience. When experiences confirm expectations, individuals attribute experienced effects to the cues that predicted them. When they do not, however, experiences are more likely to be judged to be anomalies or 'flukes', or the result of lapses in attention, and discounted. More generally, confirmation biases may reflect a broader principle of regularizing predictions by down-weighting outliers, which is the basis of robust statistics[74]. This principle can in some cases improve perceptual inference by making our estimates resistant to outliers[75,76]; hence it can be seen as a functionally adaptive strength. However, when initial beliefs are no longer valid, for example because of a change in the environment, this principle also impedes belief updating, a vulnerability that goes along with its advantages.

Though our data suggest that cues modulate the brain mechanisms underlying pain perception, they do not exclude the possibility that expectations also influence later, post-perceptual processes. That the NPS response partially mediated the effect of expectations on pain ratings suggests that part of this effect is due to modulation of relatively early pain signalling, while another part reflects effects at a later stage (such as evaluation or response). Studies using signal-detection theory have shown that placebo treatments can increase observers' pain criterion values[77,78], implying an effect at the decision or response level. Likewise, a recent study applying evidence-accumulation models of perceptual decision-making to a pain-discrimination task suggested that pain-predictive cues introduce response bias into the decision process[79]. When combining these findings with findings that expectations modify pain-related activity in the spinal cord[33,34], it seems that expectations can affect multiple stages of pain processing and evaluation. The relative strength of early and late expectancy effects probably varies across people and situations, and may depend on the magnitude of expectancy violations. Further specification of the pain-processing stages that are modified by expectations is an important objective for future research.

How prior beliefs affect perception and learning may strongly depend also on how those beliefs were acquired. In our studies, cue–pain associations were established using a learning procedure in which cues were repeatedly paired with pictures of thermometers displaying low or high heat levels. This procedure differs from conventional classical conditioning as it did not involve primary affective outcomes, and because expectations and pain were reported on each trial. The reinforcement with pictures of thermometers instead of painful heat, which we call 'symbolic conditioning', is closely aligned with sensory preconditioning paradigms in animals[80,81] and humans[82], and this type of learning may involve interactions between the hippocampus (traditionally associated with explicit memory) and other value-based learning systems[83]. In a recent study we found only weak effects of a standard classical conditioning procedure on the NPS[70], suggesting that symbolic, conceptual learning procedures may have unique effects. The act of explicitly rating expectations may have important effects as well[84]. In support of this view, recent work suggests that when threat is engaged by explicit, instruction-driven mechanisms, it produces physiological threat responses via pathways that bypass the amygdala[85]. It is possible that the expectancy ratings in our paradigm promoted the persistent explicit recall of the initial pain expectations, preventing participants from forgetting these initial expectations. Our Bayesian model captures such 'resistant priors' by the drift of expectations toward their initial values after updating, and our extended reinforcement learning model by a persistent effect of initial beliefs on pain. Whether and how the resistance to extinction depends on the specific task structure, such as the presence of explicit expectancy ratings, remains to be tested in future studies.

In addition to perception modulation, confirmation biases in learning appear to be critical for explaining the persistent influences of cues we observed. The reinforcement learning model that included confirmation bias in learning updates outperformed all other models, providing evidence for a contribution of biased learning over and above direct resistant-prior effects. However, we found substantial individual differences in the degree to which participants showed a confirmation bias in learning. Furthermore, individual differences in confirmation bias in learning rate predicted the degree to which participants showed a confirmation bias in the updating of pain-anticipatory activity in several brain regions. When worse-than-expected pain was preceded by a high- relative to a low-pain cue, participants with stronger confirmation biases on learning rate also showed a greater increase in subsequent pain-anticipatory activity in the anterior midcingulate cortex, sensorimotor cortex, mid-insula, posterior striatum, periaqueductal gray and pons. These regions have previously been associated with anticipatory anxiety[86], pain prediction errors and aversive action policy[87], and cues predictive of high pain[25,88]. The ventromedial PFC showed the opposite correlation: participants with stronger confirmation biases showed greater reductions in anticipatory ventromedial PFC activity following worse-than-expected pain preceded by high- relative to low-pain cues, and greater increases in anticipatory ventromedial PFC activity following less-than-expected pain preceded by low- relative to high-pain cues. Activity in the ventromedial PFC tracked appetitive value in many studies[89–91], and has been shown to negatively track pain expectations[87]. Thus, our results suggest that for high confirmation-bias learners, combining high pain expectancy with worse-than-expected outcomes activates systems associated with anticipatory anxiety, threat, and negative affective value. However, as these results were based on differences between single-trial fMRI activation maps—with low signal-to-noise ratio—and on a relatively small number of participants, they must be considered preliminary, awaiting replication.

The effects of expectations on pain and learning circuits demonstrated in our studies could be applied to the study of chronic pain. Self-reinforcing expectancy effects may facilitate the transition from acute to chronic pain, which is all too common after surgery or injury[92,93]. Indeed, the inability to extinguish pain memories has been proposed as a defining aspect of chronic pain[94,95]. Related to this, previous treatment experiences predict people's analgesic response to subsequent placebo treatments[96–98] and to active medical treatments[99], suggesting that positive and negative beliefs are important determinants of pain responses. Furthermore, a few studies have demonstrated impaired cue–pain contingency learning in chronic-pain patients, with a particular impairment in safety learning[100,101]. If chronic-pain patients expect pain in more situations than healthy people, these deficits may reflect reduced learning from belief-disconfirming (no pain), relative to belief-confirming (pain) outcomes; hence it is possible that many pain patients have a high confirmation-bias phenotype. The interactions between perceptual and learning processes in the development of chronic pain remain to be clarified[102]. One promising approach would be to quantify effects like those identified in our studies in people at risk for developing chronic pain.

Effects of expectations on perception and learning may also contribute to psychological disorders that are characterized by persistent negative expectations and beliefs, such as anxiety and depression. Recent studies showed that more anxious people are less able to flexibly adapt their learning rate in a dynamic aversive learning task[103], and that—unlike healthy people—socially anxious people do not show a positivity bias when updating self-related information based on social feedback[104]. Whether these findings reflect a confirmation-bias mechanism similar to that observed in our studies—for example, the down-weighting of evidence that does not confirm previously learned associations or beliefs—is currently unknown. Another open question concerns the potential contribution of fluctuations in mood to our findings. In the reward domain, expectations and prediction errors have been shown to affect people's self-reported mood on a trial-to-trial basis[105]. Furthermore, mood can bias how people perceive and learn from rewards, especially in people with high mood instability[106]. Future research is needed to further investigate the interactions between expectations, mood and affective learning.

A limitation of our studies is that prediction errors were inconsistent with the initial cue–pain associations on most of the trials. According to our models, a larger proportion of 'cue-consistent' prediction errors should produce even stronger persistence of cue-based expectations and pain modulation over time. Future studies could increase the proportion of consistent prediction errors, by using more extreme noxious-stimulus intensities, to test this hypothesis. In addition, it is unknown to what degree our expectation-updating results reflect explicit versus implicit learning

processes. We previously demonstrated that self-reported expectations and autonomic anticipatory responses have opposite effects on autonomic pain responses[27]. Thus, it is possible that there are additional learning processes that also influence pain and operate in parallel to the mechanisms revealed here. Finally, although our results provide strong evidence that the cue effects are highly resistant to extinction, we cannot be sure whether our cue effects will never extinguish, or whether they would eventually extinguish but at a much slower rate than would be expected in the absence of confirmation bias. Studies using a larger number of test trials are needed to differentiate between these possibilities.

To conclude, our findings demonstrate that initial beliefs can become self-fulfilling prophecies through their combined effects on perceptual and learning processes. Future studies may examine the generalizability of our findings to self-reinforcing phenomena in other domains, and the clinical relevance of individual differences in the effects revealed in our studies. These insights can help understand—and may eventually help to counteract—self-fulfilling phenomena in various domains of human behaviour.

Methods

Participants. Thirty participants took part in Study 1 (a behavioural study, previously published by ref. [27]), and thirty-four participants took part in Study 2 (an fMRI study). We chose the sample sizes based on large effect sizes in previous cue-based pain modulation studies conducted in our laboratory[25]. Sample sizes of about 30 provide approximately 80% power to detect an effect size of Cohen's $d = 0.54$ or larger, which is near the lower bound of what we would expect for the predictive-cue effects on pain responses. All participants were healthy and reported no history of psychiatric, neurological or pain disorders, and no current pain. All participants gave informed consent and the experiment was approved by the institutional review board of the University of Colorado Boulder. Participants in Study 1 received US$12 per hour, and participants in Study 2 received US$24 per hour for their participation.

In Study 1, one participant decided to stop before the end of the experiment because she found the heat too painful, and one participant had to be excluded because of thermode failure. Thus, the final dataset for Study 1 was based on twenty-eight participants (mean age = 25, range = 18–55 years; 25% female). In Study 2 (mean age = 25, range = 18–54 years; 50% female), one participant misunderstood the expected-pain rating procedure, and was excluded from all analyses involving pain expectations.

Experimental task. In both studies, we used a previously established paradigm consisting of a learning phase followed by a test phase (Fig. 1a). Before starting the task, we instructed participants that they would learn associations between specific shapes and specific heat levels. We also instructed them that the heat levels would be displayed on a thermometer during the first part of the experiment and applied to their skin during the second part of the experiment. In Study 2, the learning and test phases were both performed within the MRI scanner.

Learning phase. Each trial started with the 2-s presentation of a visual cue (a geometric shape). Following the cue, participants indicated which heat level they expected on a 100-unit vertical visual analogue scale (VAS) with lower and upper anchors of 'baseline skin temperature' and 'extremely hot', respectively, which was displayed alongside a picture of a blank thermometer. A few seconds later, the thermometer picture reappeared for 3 s, this time indicating a specific symbolic heat level. Some cues (low-pain cues) were always followed by low heat levels (32–53% of the thermometer scale) and other cues (high-pain cues) were always followed by high heat levels (73–93% of the thermometer scale). The learning phase of Study 1 included three low- and three high-pain cues, and participants completed 20 trials for each cue, in pseudorandom order stratified across time, with the constraint that each block of six trials contained one trial in each cell of the six cue conditions. The learning phase of Study 2 included two low- and two high-pain cues, and participants completed 30 trials for each cue, in pseudorandom order stratified across time, with the constraint that each block of four trials contained one trial in each cell of the four cue conditions.

Test phase. Each test trial started with the 2-s presentation of one of the cues from the preceding learning phase, or a novel cue (see below). After this, participants indicated how much pain they expected on that trial, on a horizontal 100-unit VAS with anchors of 'no pain' and 'worst-imaginable pain'. A few seconds later, a noxious heat stimulus was applied to participants' left inner forearm (Study 1) or lower leg (Study 2), using a 27-mm diameter contact heat-evoked potential stimulator thermode (ramp rate = 40 °C s^{-1}; 1 s at peak temperature; peak temperature = 47 °C or 48 °C in Study 1, and 48 °C or 49 °C in Study 2). Between stimuli, the thermode maintained a baseline temperature of 32 °C. Unbeknownst to the participants, all cues were followed by each of two temperatures on 50% of the trials each, that is, heat intensity was unrelated to the preceding cue (Fig. 1a). Several seconds after heat offset, the VAS re-appeared on the screen and participants rated how much pain they had experienced on that trial. Then, after a variable inter-trial interval, during which an empty screen was displayed, the next trial started. Trial procedures were identical in the two studies, expect for the inter-stimulus intervals, which were somewhat longer in Study 2 (Fig. 1b).

We introduced two novel cues in the test phase of Study 1: the words LOW and HIGH, which were instructed to precede low- and high-intensity heat, respectively. In the test phase of Study 2, we introduced three additional cues: The letters L and H, which were instructed to precede low and high heat levels, respectively, and a novel geometric shape ('neutral' cue). As previously reported[27], the LOW/L versus HIGH/H cues affected pain responses in a way similar to the way the geometric shapes that had been paired with low versus high heat representations during the learning phase. Therefore, we refer to all these cues as low- versus high-pain cues, and report results based on data pooled across these cue types.

Participants completed ten test trials for each cue, that is, a total of 80 and 70 trials in Study 1 and 2, respectively, in pseudorandom order stratified across time: in Study 1 each block of eight trials contained one trial in each cell of the eight cue conditions, and in Study 2 each block of seven trials contained one trial in each cell of the seven cue conditions. In Study 1, the test phase was divided in 5 runs of 16 trials, between which we moved the thermode to a new site on the forearm. In Study 2, the test phase was divided in 5 runs of 14 trials, between which we moved the thermode to a new site on the lower leg. Before starting each new run, we applied the highest of the two temperatures once to the to-be-stimulated skin site, to reduce the impact of site-specific habituation effects[107].

Regression analyses. We conducted multi-level regression analyses on the single-trial behavioural data and, in Study 2, NPS responses (see fMRI analysis) during the test phase, using the Multilevel Mediation toolbox (http://wagerlab.colorado.edu/tools)[25,108,109]. We tested the significance of all effects using a bootstrap procedure (10,000 bootstrap samples). All tests were 2-tailed.

We tested the effect of cue type (high- versus low-pain cue) on pain ratings and NPS responses, in two separate regression models. We also modelled the effects of heat intensity (temperature) and the Temperature × Cue type interaction. To test whether the effect of cue type extinguished over time, we also modelled the linear and quadratic effects of time (trial number) and their interactions with cue type. In addition, we modelled the linear and quadratic effects of site-specific repetition to account for pain-adaptation effects due to repeated stimulation[107].

To examine evidence for reciprocal effects of expectations and pain responses on one another, we conducted two additional sets of regression analyses. First, we tested the effect of expected-pain rating on subsequent pain ratings and NPS responses, in two separate regression models. Second, we tested the effects of pain and NPS responses on subsequent expectations, also in two separate regression models. In all four analyses, we also modelled the effects of cue type (low, high and neutral cues were coded as −1, 1 and 0, respectively) and heat intensity.

Data collection and analysis were not performed blind to the conditions of the experiments. However, we do not believe this increased the risk of bias here because all participants performed experimental conditions in computer-generated sequences, and comparisons in analyses were specified in advance.

Confirmation bias analyses. To examine evidence for a confirmation bias in expectation updating, we conducted a regression analysis on estimated trial-specific learning rate (and a parallel analysis on raw expectation updates) with regressors coding for cue type, prediction-error sign (higher- versus lower-than-expected pain), their interaction and temperature. A confirmation bias should manifest as an interaction between cue type and prediction-error sign on estimated learning rate and absolute expectation-update size. Two participants in Study 1 and 12 participants in Study 2 never reported higher-than-expected pain following high-pain cues; these participants were excluded from these analyses. In an additional analysis, we used a continuous regressor for prediction error magnitude instead of prediction-error sign, allowing inclusion of all participants.

We conducted similar confirmation bias analyses on the updating of pain-anticipatory brain activity. The dependent variable in these analyses was the change in pain-anticipatory fMRI activity across successive trials on which the same cue was presented (see fMRI analysis). As aversive and appetitive pain prediction errors (higher- and lower-than-expected pain, respectively) may affect subsequent anticipatory activity in different brain circuits, we separately tested for cue effects on anticipatory activity updating following these two types of prediction errors.

Computational models. *Model 1.* Model 1 is a process-level model built in the framework of reinforcement learning. This model contains mechanisms for perceptual inference (equation (1)) and learning (equations (2) and (4)), as well as a confirmation-bias mechanism through which initial beliefs bias learning (equation (5)). We describe each of these three components of the model in turn.

Perceptual inference. The model infers the level of perceived pain on trial t, P_t, based on the current noxious input, N_t, combined with the current cue-based

pain expectation, $E_{c,t}$, for the current cue c. Specifically, P_t is a weighted linear combination of N_t and $E_{c,t}$:

$$P_t = (1-\gamma)N_t + \gamma E_{c,t} \quad (1)$$

Parameter γ ($0 \le \gamma \le 1$) controls the relative weighting of N_t and $E_{c,t}$ such that the impact of expectations on pain increases with increasing γ.

Learning. The model's learning mechanism governs the updating of cue-based expectations in response to new pain outcomes. On each trial, the discrepancy between P_t and $E_{c,t}$ elicits a prediction error, δ_t:

$$\delta_t = P_t - E_{c,t} \quad (2)$$

This prediction error triggers the updating of the pain expectation for cue c, according to a standard reinforcement-learning algorithm ('delta rule' learning)[32]:

$$E_{c,t+1} = E_{c,t} + \alpha\delta_t \quad (3)$$

Note that, given equation (2), the expectation-updating process can also be written as:

$$E_{c,t+1} = \alpha P_t + (1-\alpha)E_{c,t} \quad (4)$$

which shows that the new expectation is a weighted average of the most recent pain experience and the old expectation. The amount of updating is determined by the learning-rate parameter α ($0 \le \alpha \le 1$): Higher values of α result in stronger updating.

Confirmation bias. Model 1 implements a confirmation bias in learning by specifying α on each trial depending on whether the sign of δ_t is consistent or inconsistent with the initial low or high pain association of the current cue:

$$\alpha_t = \begin{cases} \alpha_c \text{ if } \left(\delta_t > 0 \text{ and } c = c_{\text{high}}\right) \text{ or } \left(\delta_t < 0 \text{ and } c = c_{\text{low}}\right) \\ \alpha_i \text{ if } \left(\delta_t > 0 \text{ and } c = c_{\text{low}}\right) \text{ or } \left(\delta_t < 0 \text{ and } c = c_{\text{high}}\right) \end{cases} \quad (5)$$

Here c_{high} and c_{low} refer to cues that were initially associated with high and low relative pain values, respectively. If α_c is higher than α_i, the model produces a confirmation bias. Note that the neutral cues (present only in Study 2) had not been included in the learning phase, and hence were not expected to have a low or high pain association at the start of the test phase. Therefore, we excluded the neutral-cue trials from all model fitting procedures.

Thus, Model 1 has three free parameters: γ, α_c and α_i. To optimize fits, we initialized the cue-based expectations ($E_{c,1}$) to the first expected-pain rating for cue c in the test phase, separately for each participant. In addition, we modelled the noxious input on each trial (N_t) as the participant's average pain rating for the temperature presented on that trial, such that N_t and P_t are represented on the same scale.

Persistent effects of initial beliefs. In the extended model (Model 1$^+$) that allowed pain perception to be affected by the initial beliefs acquired during the learning phase, equation (1) was replaced by

$$P_t = (1-\lambda-\gamma)N_t + \gamma E_{c,t} + \lambda M_c \quad (6)$$

where M_c is the initial value associated with cue c, acquired during the learning phase. We set M_c to the first expected-pain rating for cue c in the test phase (before the first heat stimulus for that cue had been received), separately for each participant. The additional parameter λ controls the relative weighting of this initial belief, with $0 \le \lambda \le 1-\gamma$.

Model 2. Model 2 represents pain perception and learning as Bayesian inferences. Below, we first describe the generative assumptions of this model (equations (7) to (10)), followed by its perceptual inference (equations (11) to (14)) and learning (equations (15) to (18)) components, and a mechanism to capture experience-resistant expectations (equations (19) to (22)).

Generative model. As a Bayesian model, Model 2 attributes to the subject a structured belief, that is, a generative model, regarding the dynamics and statistics of the task environment. The subject is assumed to engage in optimal Bayesian inference with respect to this generative model, based on the noxious input (that is, heat stimuli). The subject's posterior beliefs then determine his or her responses (that is, reported expectations and pain ratings).

The generative model is assumed to contain two sets of latent variables. The first, π_t, represents the true level of potential harm on each trial, in other words the objectively correct level of pain that the subject should feel. The second, $\mu_{c,t}$, represents the mean level of potential harm signalled by each cue c at any time, which is the average value for π_t if c is the cue presented on trial t. The subject's belief about $\mu_{c,t}$ at the start of trial t (after the cue is presented and before the heat stimulus) determines his or her expected pain intensity, E_t. The subject's posterior belief about π_t after the heat stimulus has been presented determines the pain report, P_t.

The subject is assumed to treat the observed noxious input, N_t, as an imperfect indicator of π_t, with variance σ_ε^2:

$$N_t \approx \mathcal{N}(\pi_t, \sigma_\varepsilon^2) \quad (7)$$

Here $\mathcal{N}(\pi_t, \sigma_\varepsilon^2)$ indicates a Gaussian distribution with mean π_t and variance σ_ε^2. This relationship defines a likelihood for the observed value of N_t, which can be combined with the prior based on μ_c to derive a posterior for π_t using Bayes' rule (perceptual inference, elaborated below).

Model 2 assumes that participants track beliefs about μ_c for each cue (learning) using a Kalman filter, a well studied model of Bayesian learning that has been used to model learning of reward magnitudes in previous studies[110,111]. The Kalman filter assumes a generative model in which π_t is sampled from a Gaussian distribution centred on $\mu_{c,t}$ with a fixed standard deviation, σ_ψ^2:

$$\pi_t \approx \mathcal{N}(\mu_{c,t}, \sigma_\psi^2) \quad (8)$$

The Kalman filter also assumes that $\mu_{c,t}$ varies over time (hence the additional index t), according to a Gaussian random walk with mean step size 0 and standard deviation σ_η^2:

$$\mu_{c,t+1} \approx \mathcal{N}(\mu_{c,t}, \sigma_\eta^2) \quad (9)$$

Extending the standard Kalman filter, we assume the subject maintains a conjugate iterative prior on $\mu_{c,t}$ conditioned on all previous observations:

$$\mu_{c,t}|\mathbf{N}_{t-1} \approx \mathcal{N}(m_{c,t}, s_{c,t}^2) \quad (10)$$

where $N_{t-1} = (N_1, \ldots, N_{t-1})$ is the sequence of all prior inputs, and $m_{c,t}$ and $s_{c,t}$ are values that the subject tracks as part of learning. That is, $m_{c,t}$ and $s_{c,t}$ characterize the subject's state of knowledge regarding $\mu_{c,t}$ at the beginning of trial t. Both levels of inference in the model are based on the iterative prior in equation (10).

Perceptual inference. The iterative prior on $\mu_{c,t}$ also yields a prior on π_t, before the heat stimulus on each trial, obtained by adding the trial-level variance σ_ψ^2:

$$\pi_t|\mathbf{N}_{t-1} \approx \mathcal{N}(m_{c,t}, s_{c,t}^2 + \sigma_\psi^2) \quad (11)$$

The mean of this prior, $m_{c,t}$, is the participant's reported expectation, E_t, at the beginning of a trial (that is, after the cue has been observed):

$$E_t = m_{c,t} \quad (12)$$

Once N_t is observed, it can be combined with the prior in equation (11) to derive a posterior for π_t. Using Bayes' rule, this posterior can be shown to be:

$$\pi_t|\mathbf{N}_t \approx \mathcal{N}\left(\frac{\sigma_\varepsilon^2 m_{c,t} + (\sigma_\psi^2 + s_{c,t}^2)N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_{c,t}^2}, \frac{\sigma_\varepsilon^2(\sigma_\psi^2 + s_{c,t}^2)}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_{c,t}^2}\right) \quad (13)$$

The mean of this posterior is the subject's pain report, P_t, which can also be written as:

$$P_t = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_{c,t}^2} m_{c,t} + \frac{\sigma_\psi^2 + s_{c,t}^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_{c,t}^2} N_t \quad (14)$$

Thus, as in Model 1, reported pain is a weighted average of the expectation and noxious input. The weights in this case are determined by the uncertainties in the two sources of information (the prior expectation and the observed input), each weighted in proportion to its precision (inverse variance).

Learning. Once the posterior belief for π_t is calculated, it is used to calculate the posterior distribution for $\mu_{c,t}$. As with equation (13), the mean of this posterior can be shown to be a precision-weighted average of the prior belief (equation (10)) and the new observation (N_t):

$$\mu_{c,t}|\mathbf{N}_t \approx \mathcal{N}\left(\frac{(\sigma_\varepsilon^2 + \sigma_\psi^2)m_{c,t} + s_{c,t}^2 N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_{c,t}^2}, \frac{(\sigma_\varepsilon^2 + \sigma_\psi^2)s_{c,t}^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_{c,t}^2}\right) \quad (15)$$

The prior on the next trial is then obtained by adding the variance of the random walk:

$$\mu_{c,t+1}|\mathbf{N}_t \approx \mathcal{N}\left(\frac{(\sigma_\varepsilon^2 + \sigma_\psi^2)m_{c,t} + s_{c,t}^2 N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_{c,t}^2}, \frac{(\sigma_\varepsilon^2 + \sigma_\psi^2)s_{c,t}^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_{c,t}^2} + \sigma_\eta^2\right) \quad (16)$$

By definition, the mean and variance of this iterative prior equal $m_{c,t+1}$ and $s^2_{c,t+1}$, respectively. This recursive relation yields update equations for m:

$$m_{c,t+1} = \frac{\sigma_\varepsilon^2 + \sigma_\psi^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} m_{c,t} + \frac{s^2_{c,t}}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} N_t \quad (17)$$

and for s:

$$s^2_{c,t+1} = \frac{(\sigma_\varepsilon^2 + \sigma_\psi^2) s^2_{c,t}}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} + \sigma_\eta^2 \quad (18)$$

Equation (17) shows that the new expectation is a weighted average of the most recent noxious input and the old expectation, similar to Model 1.

Biased learning. To model a persistent influence of initial expectations, we put a bias into the random walk, as a spatial (that is, directional) inhomogeneity. Specifically, we included a directed component in the dynamics of the generative model, replacing equation (9) with

$$\mu_{c,t+1} \approx \mathcal{N}((1-\beta)\mu_{c,t} + \beta M_c, \sigma_\eta^2) \quad (19)$$

where M_c is the initial value associated with cue c during the learning phase. Thus, the participant's generative model assumes μ_c tends to decay toward or away from M_c, and the direction and strength of this effect are determined by parameter $\beta \in [-1, 1]$. Values of $\beta > 0$ yield a decay of μ_c toward M_c (persistent beliefs), and values of $\beta < 0$ yield a growth of μ_c away from M_c (the opposite of persistent beliefs).

This change to the random walk between trials has no impact on inference within a trial. Thus, equations (11) to (15) are unchanged. The only change is to the learning step between trials, with the iterative prior in equation (16) replaced by

$$\mu_{c,t+1} | \mathbf{N}_t \approx \mathcal{N}\left(\frac{(1-\beta)(\sigma_\varepsilon^2 + \sigma_\psi^2) m_{c,t} + (1-\beta) s^2_{c,t} N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} + \beta M_c, \frac{(1-\beta)^2 (\sigma_\varepsilon^2 + \sigma_\psi^2) s^2_{c,t}}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} + \sigma_\eta^2 \right) \quad (20)$$

The corresponding update rules (replacing equations (17) to (18)) are then

$$m_{c,t+1} = \frac{(1-\beta)(\sigma_\varepsilon^2 + \sigma_\psi^2)}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} m_{c,t} + \frac{(1-\beta) s^2_{c,t}}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} N_t + \beta M_c \quad (21)$$

and

$$s^2_{c,t+1} = \frac{(1-\beta)^2 (\sigma_\varepsilon^2 + \sigma_\psi^2) s^2_{c,t}}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} + \sigma_\eta^2 \quad (22)$$

Equation (21) shows how the update of the expectation for the next trial ($m_{c,t+1}$) is biased toward the initial association value for the current cue (M_c), provided $\beta > 0$.

In summary, Model 2 has four free parameters: the directional bias of the random walk in cue values (β), the variance of the random walk in cue values (σ_η^2), the variance in potential harm on any given trial (σ_ψ^2), and the noise variance of noxious input (σ_ε^2). For the last three parameters, only their ratios matter; hence we fixed σ_ε^2 at a value of 1 and estimated the other two. For the reduced models with σ_ε^2 fixed to zero, we fixed σ_ψ^2 at a value of 1 and estimated σ_η^2.

As in Model 1, we initialized the mean cue-based expectations (prior to trial 1; $m_{c,1}$ for each value of c) to the participant's expected-pain rating following the first appearance of each cue in the test phase, and we modelled the noxious input on each trial (N_t) as the participant's average pain rating across all trials on which that temperature was presented (to put temperature and pain on the same scale). Finally, we assumed that the initial variance of the subject's prior for $\mu_{c,1}$ (that is, the subject's uncertainty, $s^2_{c,1}$) was equal to σ_η^2 for all cues.

Comparison of Models 1 and 2. Despite their substantial differences in framing, the reinforcement learning and Bayesian models presented here can be seen as embodying a shared set of theoretical principles. The main difference is that the former is cast at an algorithmic or mechanistic level, whereas the latter is cast at a computational level, in terms of optimal statistical inference. Specifically, adjusting expectations toward new observations as in delta-rule learning (equations (2) to (4)) can be seen as an algorithmic solution to tracking a moving target as in the Kalman filter (equation (15))[1,57,112–114]. Likewise, perception as a weighted combination of sensory input and expectations (equation (1)) enacts a Bayesian integration of new data and prior beliefs (equations (13) and (14)), or more generally of different sources of information in proportion to their precision values (inverse variances)[115–118].

To see these connections formally, define a trial-specific weighting parameter in Model 2:

$$\gamma_t = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s^2_{c,t}} \quad (23)$$

Then from equations (12) and (14), the perceptual inference step in Model 2 yields a mean pain report of:

$$P_t = (1-\gamma_t) N_t + \gamma_t E_{c,t} \quad (24)$$

in agreement with the perceptual inference step in Model 1 (equation (1)). Thus perceptual inference works in the same way in both models, except that Model 2 determines the weighting parameter (γ) rationally and dynamically, based on current levels of uncertainty in the current input N_t (σ_ε^2) and in the prior for π_t ($\sigma_\psi^2 + s^2_{c,t}$) as indicators of the level of potential harm (that is, of the true value of π_t).

The learning steps of the models can be similarly linked. First consider Model 2 without biased learning ($\beta = 0$), and define a trial-specific learning rate:

$$\alpha_t = \frac{s^2_{c,t}}{\sigma_\psi^2 + s^2_{c,t}} \quad (25)$$

From equations (17) and (14), the learning step in Model 2 yields a mean expectation on the next trial of:

$$E_{c,t+1} = \alpha_t P_t + (1-\alpha_t) E_{c,t} \quad (26)$$

in agreement with the learning update in Model 1 (equation (4)). Thus learning works in the same way in both models, except that Model 2 determines the learning rate (α) rationally and dynamically, based on current levels of uncertainty in π_t (σ_ψ^2) and in the prior for $\mu_{c,t}$ ($s^2_{c,t}$) as indicators of the objective cue-pain association (that is, of the true value of $\mu_{c,t}$).

With bias in the random walk (governed by β), the learning update in Model 2 (equation (21)) becomes:

$$E_{c,t+1} = \alpha_t (1-\beta) P_t + (1-\alpha_t)(1-\beta) E_{c,t} + \beta M_c \quad (27)$$

Thus the initial belief directly contributes to the learning update, similar to its role in perceptual inference in Model 1$^+$. A second interpretation of Model 2's learning update (also equivalent to equation (21)) that is closer to that of the primary Model 1 can be obtained by defining an error-dependent learning rate:

$$\tilde{\alpha}_t = (1-\beta)\alpha_t + \beta \frac{M_c - E_{c,t}}{P_t - E_{c,t}} \quad (28)$$

Using this definition, equation (21) can be rewritten as:

$$E_{c,t+1} = \tilde{\alpha}_t P_t + (1-\tilde{\alpha}_t) E_{c,t} \quad (29)$$

paralleling equation (4). Moreover, similar to α_c and α_i in Model 1 (see equation (5)), a positive value of β implies that $\tilde{\alpha}_t$ is larger when the prediction error moves the expectation toward the initial association (that is, $P_t - E_{c,t}$ and $M_c - E_{c,t}$ have the same sign) and smaller otherwise. The main difference is that Model 1 implements biased learning by assuming two fixed learning rates for cue-consistent and cue-inconsistent prediction errors, whereas (under the interpretation of equations (28) and (29)), Model 2 implements biased learning by allowing learning rate to vary continuously as a function of the relationships among E_c, M_c and P_t.

Model fitting. To estimate the model parameters and assess relative predictive quality of each of the models, we implemented the models in Stan[119]. Stan enables Bayesian inference through Markov chain Monte Carlo (MCMC) sampling from the posterior distribution over model parameters. We note that the choice to estimate the models using Bayesian methods is unrelated to the assumption in Model 2 that subjects learn using Bayesian inference. Both Model 1 and Model 2 were fitted hierarchically, such that the model parameters of each participant are sampled from a group-level distribution. In this way, the information in the individual data is aggregated, while still respecting individual differences[120]. In the Bayesian framework, this means that each individual-level parameter is assigned a group-level prior distribution, whose distribution parameters (hyperparameters) are assigned prior distributions (hyperpriors).

Group-level distributions for α_c, α_i and γ in Model 1 and parameter β in Model 2 were assumed to be beta distributions. The hyperparameters defining each of these beta distributions were assigned uniform hyperpriors on the interval [0,10]. Parameters σ_η^2 and σ_ψ^2 in Model 2 were assumed to have half-Cauchy distributions. In both models, subjects' expectation and pain ratings were assumed to be normally distributed around the model's point predictions. The variances of both of these normal distributions (error variances) were each assumed to have a half-Cauchy group-level distribution. We used uniform

distributions on the interval [0,10] as hyperpriors for the scale parameters governing the half-Cauchy distributions. These priors and hyperpriors were chosen for their uninformative nature[121].

For both models, the number of burn-in iterations was set at 5,000, the number of posterior samples taken was set at 10,000, and four MCMC chains were run with overdispersed starting values. All chains showed proper convergence, as assessed by visual inspection, and all Rhat values were less than 1.1.

For model comparison, we used the bridge sampling algorithm[122–124] in R to obtain Bayes factors[125]. The Bayes factor provides a relative metric between two models' predictive performances. A Bayes factor of 1 corresponds to both models performing equally well. A Bayes factor of 5, for example, can be interpreted as the data being five times as likely under the first model compared to the second. Furthermore, the Bayes factor automatically incorporates parsimony, so that less complex models are preferred over more complex models when they would fit equally well under maximum likelihood.

FMRI acquisition and preprocessing. *Imaging acquisition.* In Study 2, we acquired whole-brain fMRI data on a Siemens 3T Trio scanner at the Center for Innovation and Creativity in Boulder. Structural images were acquired using high-resolution T1 spoiled gradient recall images for anatomical localization and warping to a standard space. Functional images were acquired with an echo-planar imaging sequence (repetition time (TR) = 1300 ms, echo time (TE) = 25 ms, field of view = 220 mm, 3.4 × 3.4 × 3.0 mm voxels, 26 slices, parallel imaging, SENSE factor 2). Each test-phase run lasted 430 s (331 TRs). Visual stimuli were presented via a mirror attached to the head coil, and ratings were made using a track ball.

Preprocessing. Prior to preprocessing, global outlier time points (that is, 'spikes' in the BOLD signal) were identified by computing both the mean and the standard deviation (across voxels) of values for each image for all slices. Mahalanobis distances for the matrix of slice-wise mean and standard deviation values (concatenated) × functional volumes (time) were computed, and any values with a significant $\chi 2$ value (corrected for multiple comparisons based on the more stringent of either false discovery rate or Bonferroni methods) were considered outliers. On average 4.6% of images were outliers (s.d. = 1.9). The output of this procedure was later used as a covariate of noninterest in the first-level models.

Functional images were slice-acquisition-timing and motion corrected using SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK). Structural T1-weighted images were coregistered to the first functional image for each subject using an iterative procedure of automated registration using mutual information coregistration in SPM8 and manual adjustment of the automated algorithm's starting point until the automated procedure provided satisfactory alignment. Structural images were normalized to Montreal Neurological Institute (MNI) space using SPM8, interpolated to 2 × 2 × 2 mm voxels, and smoothed using a 6mm full-width at half maximum Gaussian kernel. We discarded the first 6 volumes of each run, and then concatenated the 5 test-phase runs for each participant. A high-pass filter of 180 seconds was used.

fMRI analysis. We conducted first-level (individual participants) general linear model analyses in SPM8, employing the single trial, or "single-epoch", design and analysis approach[49,63,64]. To estimate single-trial pain responses, we constructed a general linear model design matrix with separate regressors for the 1.8-s heat-application period in each trial. We also modelled periods of cue presentation (2 s), expected-pain rating (VAS onset to response), pain anticipation (expected-pain rating response to heat onset), and pain rating (VAS onset to response). All task variables were modelled as boxcar regressors, convolved with the canonical hemodynamic response function. Other regressors of non-interest (nuisance variables) were (1) "dummy" regressors coding for each run (intercept for each but the last run); (2) linear drift across time within each run; (3) the six estimated head movement parameters (x, y, z, roll, pitch and yaw), their mean-zeroed squares, their derivatives, and squared derivatives for each run (total 24 columns per run); (4) indicator vectors for outlier time points identified based on their multivariate distance from the other images in the sample (see above); and (5) indicator vectors for the first two images in each run.

In a second general linear model we used separate regressors for each single-trial pain-anticipation period. The only difference between this general linear model and the one reported above is that the pain-anticipation periods, instead of the heat-application periods, were modelled on a single-trial basis.

One important consideration in single-trial analysis is that trial estimates can be strongly affected by acquisition artefacts that occur during that trial (such as sudden motion or scanner pulse artefacts). For this reason, we calculated trial-by-trial variance inflation factors (a measure of design-induced uncertainty, in this case due to collinearity with nuisance regressors), and trials with variance inflation factors that exceeded 2.5 were excluded from the analyses. The average number of excluded trials was 2.3 per participant (s.d. = 2.2).

NPS analyses. We calculated the strength of expression of the NPS pattern for each single-trial heat activation map, by taking the dot product of the vectorized activation image (β_{map}) with the NPS pattern weights ($\mathbf{w}_{map}$), yielding a continuous scalar value. We used the single-trial NPS responses in our regression analyses.

Anticipatory activity updating analysis. To examine evidence for a confirmation bias in the updating of pain-anticipatory brain activity, we computed the change in pain-anticipatory activity across successive presentations of a given cue. Specifically, we created 'delta' (or 'update') images by subtracting the single-trial activation map on trial t from the single-trial activation map for the next trial on which the same cue c was presented, that is, activation $map_{c,t+1}$ − activation $map_{c,t}$. We then separately computed each participant's mean update image following aversive prediction errors on low-pain-cue trials, aversive prediction errors on high-pain-cue trials, appetitive prediction errors on low-pain-cue trials, and appetitive prediction errors on high-pain-cue trials. Finally, we computed the following two contrasts: (1) the update of anticipatory activation following aversive prediction errors on high-pain-cue > low-pain-cue trials; and (2) the update of anticipatory activation following appetitive prediction errors on low-pain-cue > high-pain-cue trials.

We conducted a second-level (group) analysis on each of these two contrasts using robust regression, which minimizes the influence of outliers[126], and included a regressor coding for individual differences in estimated learning rate ($\hat{\alpha}$) between high-pain-cue and low-pain-cue trials (either for aversive or appetitive prediction errors, for the corresponding contrast). We applied a gray-matter mask, and used FDR to correct for multiple comparisons in whole-brain voxel-wise analyses.

Activation clusters, reported in the Supplementary Tables, were defined as FDR-corrected voxels ($q < 0.05$) contiguous with voxels at uncorrected $P < 0.001$ and $P < 0.01$. In the figures in the main manuscript, we show FDR-corrected voxels ($q < 0.05$) contiguous with voxels at uncorrected $P < 0.01$ and $P < 0.05$, for display purposes.

Code availability. Our modelling code (all versions of our computational models) is available through the Open Science Framework repository, https://osf.io/bqkz3/. The code we used to conduct all other analyses (multilevel regression, mediation and fMRI analyses) is available via https://github.com/canlab.

Data availability

The single-trial behavioural and NPS data, which are needed to reproduce all behavioural and NPS analyses in the paper, are available through the Open Science Framework repository, https://osf.io/bqkz3/. The fMRI data, which are needed to reproduce the analyses on anticipatory brain activity, are available from the corresponding author upon request.

Received: 7 March 2017; Accepted: 19 September 2018;
Published online: 29 October 2018

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, 1998).
2. Pavlov, I. P. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (Dover Publications, New York, 1960).
3. Benedetti, F. Placebo effects: from the neurobiological paradigm to translational implications. *Neuron* **84**, 623–637 (2014).
4. Benedetti, F., Carlino, E. & Pollo, A. How placebos change the patient's brain. *Neuropsychopharmacology* **36**, 339–354 (2011).
5. Colloca, L. & Benedetti, F. Placebos and painkillers: is mind as real as matter? *Nat. Rev. Neurosci.* **6**, 545–552 (2005).
6. Wager, T. D. & Atlas, L. Y. The neuroscience of placebo effects: connecting context, learning and health. *Nat. Rev. Neurosci.* **16**, 403–418 (2015).
7. Oken, B. S. Placebo effects: clinical aspects and neurobiology. *Brain* **131**, 2812–2823 (2008).
8. Price, D. D., Finniss, D. G. & Benedetti, F. A comprehensive review of the placebo effect: recent advances and current thought. *Annu. Rev. Psychol.* **59**, 565–590 (2008).
9. Walsh, B. T., Seidman, S. N., Sysko, R. & Gould, M. Placebo response in studies of major depression: variable, substantial, and growing. *J. Am. Med. Assoc.* **287**, 1840–1847 (2002).
10. Sterzer, P., Frith, C. & Petrovic, P. Believing is seeing: expectations alter visual awareness. *Curr. Biol.* **18**, R697–R698 (2008).
11. Summerfield, C. & Egner, T. Expectation (and attention) in visual cognition. *Trends. Cogn. Sci.* **13**, 403–409 (2009).
12. Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14**, 350–363 (2013).
13. Nitschke, J. B. et al. Altering expectancy dampens neural response to aversive taste in primary taste cortex. *Nat. Neurosci.* **9**, 435–442 (2006).
14. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
15. Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B. Biol. Sci.* **216**, 427–459 (1982).
16. Buchel, C., Geuter, S., Sprenger, C. & Eippert, F. Placebo analgesia: a predictive coding perspective. *Neuron* **81**, 1223–1239 (2014).

17. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. Lond. B* **364**, 1211–1221 (2009).
18. Friston, K. A theory of cortical responses. *Phil. Trans. R. Soc. Lond. B* **360**, 815–836 (2005).
19. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
20. Merton, R. K. The self-fulfilling prophecy. *Antioch Rev.* **8**, 193–210 (1948).
21. Wager, T. D., Scott, D. J. & Zubieta, J. K. Placebo effects on human mu-opioid activity during pain. *Proc. Natl Acad. Sci. USA* **104**, 11056–11061 (2007).
22. Wiech, K. Deconstructing the sensation of pain: the influence of cognitive processes on pain perception. *Science* **354**, 584–587 (2016).
23. Atlas, L. Y. & Wager, T. D. How expectations shape pain. *Neurosci. Lett.* **520**, 140–148 (2012).
24. Montgomery, G. H. & Kirsch, I. Classical conditioning and the placebo effect. *Pain* **72**, 107–113 (1997).
25. Atlas, L. Y., Bolger, N., Lindquist, M. A. & Wager, T. D. Brain mediators of predictive cue effects on perceived pain. *J. Neurosci.* **30**, 12964–12977 (2010).
26. Colloca, L., Petrovic, P., Wager, T. D., Ingvar, M. & Benedetti, F. How the number of learning trials affects placebo and nocebo responses. *Pain* **151**, 430–439 (2010).
27. Jepma, M. & Wager, T. D. Conceptual conditioning: mechanisms mediating conditioning effects on pain. *Psychol. Sci.* **26**, 1728–1739 (2015).
28. Koban, L. & Wager, T. D. Beyond conformity: social influences on pain reports and physiology. *Emotion* **16**, 24–32 (2016).
29. Vase, L., Norskov, K. N., Petersen, G. L. & Price, D. D. Patients' direct experiences as central elements of placebo analgesia. *Phil. Trans. R. Soc. Lond. B* **366**, 1913–1921 (2011).
30. Vase, L., Robinson, M. E., Verne, G. N. & Price, D. D. Increased placebo analgesia over time in irritable bowel syndrome (IBS) patients is associated with desire and expectation but not endogenous opioid mechanisms. *Pain* **115**, 338–347 (2005).
31. Craggs, J. G., Price, D. D., Perlstein, W. M., Verne, G. N. & Robinson, M. E. The dynamic mechanisms of placebo induced analgesia: evidence of sustained and transient regional involvement. *Pain* **139**, 660–669 (2008).
32. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Research and Theory* (eds Black, A. H. & Prokasy, W. F.) 64–99 (Appleton-Century-Crofts, New York, 1972).
33. Eippert, F., Finsterbusch, J., Bingel, U. & Buchel, C. Direct evidence for spinal cord involvement in placebo analgesia. *Science* **326**, 404 (2009).
34. Geuter, S. & Buchel, C. Facilitation of pain in the human spinal cord by nocebo treatment. *J. Neurosci.* **33**, 13784–13790 (2013).
35. Plassmann, H., O'Doherty, J., Shiv, B. & Rangel, A. Marketing actions can modulate neural representations of experienced pleasantness. *Proc. Natl Acad. Sci. USA* **105**, 1050–1054 (2008).
36. Doll, B. B., Hutchison, K. E. & Frank, M. J. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J. Neurosci.* **31**, 6188–6198 (2011).
37. Doll, B. B., Jacobs, W. J., Sanfey, A. G. & Frank, M. J. Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain Res.* **1299**, 74–94 (2009).
38. Biele, G., Rieskamp, J., Krugel, L. K. & Heekeren, H. R. The neural basis of following advice. *PLoS Biol.* **9**, e1001089 (2011).
39. Li, J., Delgado, M. R. & Phelps, E. A. How instructed knowledge modulates the neural systems of reward learning. *Proc. Natl Acad. Sci. USA* **108**, 55–60 (2011).
40. Staudinger, M. R. & Buchel, C. How initial confirmatory experience potentiates the detrimental influence of bad advice. *Neuroimage* **76**, 125–133 (2013).
41. Biele, G., Rieskamp, J. & Gonzalez, R. Computational models for the combination of advice and individual learning. *Cogn. Sci.* **33**, 206–242 (2009).
42. Apkarian, A. V., Bushnell, M. C., Treede, R. D. & Zubieta, J. K. Human brain mechanisms of pain perception and regulation in health and disease. *Eur. J. Pain* **9**, 463–484 (2005).
43. Peyron, R., Laurent, B. & Garcia-Larrea, L. Functional imaging of brain responses to pain. *Neurophysiol. Clin.* **30**, 263–288 (2000).
44. Coghill, R. C. et al. Distributed processing of pain and vibration by the human brain. *J. Neurosci.* **14**, 4095–4108 (1994).
45. Rainville, P., Bushnell, M. C. & Duncan, G. H. Representation of acute and persistent pain in the human CNS: potential implications for chemical intolerance. *Ann. N. Y. Acad. Sci.* **933**, 130–141 (2001).
46. Mazzola, L., Isnard, J., Peyron, R., Guenot, M. & Mauguiere, F. Somatotopic organization of pain responses to direct electrical stimulation of the human insular cortex. *Pain* **146**, 99–104 (2009).
47. Johansen, J. P., Fields, H. L. & Manning, B. H. The affective component of pain in rodents: direct evidence for a contribution of the anterior cingulate cortex. *Proc. Natl Acad. Sci. USA* **98**, 8077–8082 (2001).
48. Johansen, J. P. & Fields, H. L. Glutamatergic activation of anterior cingulate cortex produces an aversive teaching signal. *Nat. Neurosci.* **7**, 398–403 (2004).
49. Woo, C. W., Roy, M., Buhle, J. T. & Wager, T. D. Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS Biol.* **13**, e1002036 (2015).
50. Lopez-Sola, M. et al. Towards a neurophysiological signature for fibromyalgia. *Pain* **158**, 34–47 (2017).
51. Lindquist, M. A. et al. Group-regularized individual prediction: theory and application to pain. *Neuroimage* **145**, 274–287 (2017).
52. Krishnan, A. et al. Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *eLife* **5**, e15166 (2016).
53. Wager, T. D. et al. An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
54. Tabor, A., Thacker, M. A., Moseley, G. L. & Kording, K. P. Pain: a statistical account. *PLoS Comput. Biol.* **13**, e1005142 (2017).
55. Anchisi, D. & Zanon, M. A Bayesian perspective on sensory and cognitive integration in pain perception and placebo analgesia. *PLoS ONE* **10**, e0117270 (2015).
56. Grahl, A., Onat, S. & Buchel, C. The periaqueductal gray and Bayesian integration in placebo analgesia. *eLife* **7**, e32930 (2018).
57. Dayan, P. & Kakade, S. in *Advances in Neural Information Processing Systems* Vol. 13 (eds Dietterich, T. G., Leen, T. K. & Tresp, V.) 451–457 (MIT Press, Cambridge, 2000).
58. Kalman, R. E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**, 35–45 (1960).
59. Koyama, T., McHaffie, J. G., Laurienti, P. J. & Coghill, R. C. The subjective experience of pain: where expectations become reality. *Proc. Natl Acad. Sci. USA* **102**, 12950–12955 (2005).
60. Wager, T. D., Atlas, L. Y., Leotti, L. A. & Rilling, J. K. Predicting individual differences in placebo analgesia: contributions of brain activity during anticipation and pain experience. *J. Neurosci.* **31**, 439–452 (2011).
61. Porro, C. A. et al. Does anticipation of pain affect cortical nociceptive systems? *J. Neurosci.* **22**, 3206–3214 (2002).
62. Lin, C. S., Hsieh, J. C., Yeh, T. C., Lee, S. Y. & Niddam, D. M. Functional dissociation within insular cortex: the effect of pre-stimulus anxiety on pain. *Brain Res.* **1493**, 40–47 (2013).
63. Rissman, J., Gazzaley, A. & D'Esposito, M. Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage* **23**, 752–763 (2004).
64. Mumford, J. A., Davis, T. & Poldrack, R. A. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage* **103**, 130–138 (2014).
65. Rosenthal, R. & Jacobson, L. *Pygmalion in the Classroom; Teacher Expectation and Pupils' Intellectual Development* (Holt, New York, 1968).
66. Bonte, M., Parviainen, T., Hytonen, K. & Salmelin, R. Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cereb. Cortex* **16**, 115–123 (2006).
67. Firestone, C. & Scholl, B. J. Cognition does not affect perception: evaluating the evidence for 'top-down' effects. *Behav. Brain Sci.* **39**, e229 (2016).
68. Ma, Y. et al. Serotonin transporter polymorphism alters citalopram effects on human pain responses to physical pain. *Neuroimage* **135**, 186–196 (2016).
69. Brascher, A. K., Becker, S., Hoeppli, M. E. & Schweinhardt, P. Different brain circuitries mediating controllable and uncontrollable pain. *J. Neurosci.* **36**, 5013–5025 (2016).
70. Woo, C. W. et al. Quantifying cerebral contributions to pain beyond nociception. *Nat. Commun.* **8**, 14211 (2017).
71. Becker, S., Gandhi, W., Pomares, F., Wager, T. D. & Schweinhardt, P. Orbitofrontal cortex mediates pain inhibition by monetary reward. *Soc. Cogn. Affect. Neurosci.* **12**, 651–661 (2017).
72. Jones, E. E. *Attribution: Perceiving the Causes of Behavior* (General Learning Press, Morristown, 1972).
73. Weiner, B. *An Attributional Theory of Motivation and Emotion* (Springer-Verlag, New York, 1986).
74. Huber, P. J. *Robust Statistics* (Wiley, New York, 1981).
75. Landy, M. S., Maloney, L. T., Johnston, E. B. & Young, M. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res.* **35**, 389–412 (1995).
76. de Gardelle, V. & Summerfield, C. Robust averaging during perceptual judgment. *Proc. Natl Acad. Sci. USA* **108**, 13341–13346 (2011).
77. Clark, W. C. & Yang, J. C. Acupunctural analgesia? Evaluation by signal detection theory. *Science* **184**, 1096–1098 (1974).
78. Clark, W. C. Sensory-decision theory analysis of the placebo effect on the criterion for pain and thermal sensitivity. *J. Abnorm. Psychol.* **74**, 363–371 (1969).
79. Wiech, K. et al. Influence of prior information on pain involves biased perceptual decision-making. *Curr. Biol.* **24**, R679–R681 (2014).

80. Lavin, M. J. Establishment of flavor-flavor associations using a sensory preconditioning training procedure. *Learn. Motiv.* **7**, 173–183 (1976).
81. Rizley, R. C. & Rescorla, R. A. Associations in second-order conditioning and sensory preconditioning. *J. Comp. Physiol. Psychol.* **81**, 1–11 (1972).
82. White, K. & Davey, G. C. Sensory preconditioning and UCS inflation in human 'fear' conditioning. *Behav. Res. Ther.* **27**, 161–166 (1989).
83. Wimmer, G. E. & Shohamy, D. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270–273 (2012).
84. Coppens, E., Spruyt, A., Vandenbulcke, M., Van Paesschen, W. & Vansteenwegen, D. Classically conditioned fear responses are preserved following unilateral temporal lobectomy in humans when concurrent US-expectancy ratings are used. *Neuropsychologia* **47**, 2496–2503 (2009).
85. Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D. & Phelps, E. A. Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *eLife* **5**, e15192 (2016).
86. Yang, H. et al. Striatal-limbic activation is associated with intensity of anticipatory anxiety. *Psychiat. Res.* **204**, 123–131 (2012).
87. Roy, M. et al. Representation of aversive prediction errors in the human periaqueductal gray. *Nat. Neurosci.* **17**, 1607–1612 (2014).
88. Seymour, B. et al. Temporal difference models describe higher-order learning in humans. *Nature* **429**, 664–667 (2004).
89. O'Doherty, J. P. Contributions of the ventromedial prefrontal cortex to goal-directed action selection. *Ann. N. Y. Acad. Sci.* **1239**, 118–129 (2011).
90. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).
91. Hare, T. A., Camerer, C. F. & Rangel, A. Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* **324**, 646–648 (2009).
92. Flor, H. New developments in the understanding and management of persistent pain. *Curr. Opin. Psychiatry* **25**, 109–113 (2012).
93. Soderlund, A. The role of educational and learning approaches in rehabilitation of whiplash-associated disorders in lessening the transition to chronicity. *Spine* **36**, S280–S285 (2011).
94. Mansour, A. R., Farmer, M. A., Baliki, M. N. & Apkarian, A. V. Chronic pain: the role of learning and brain plasticity. *Restor. Neurol. Neurosci.* **32**, 129–139 (2014).
95. Apkarian, A. V. Pain perception in relation to emotional learning. *Curr. Opin. Neurobiol.* **18**, 464–468 (2008).
96. Colloca, L. & Benedetti, F. How prior experience shapes placebo analgesia. *Pain* **124**, 126–133 (2006).
97. Andre-Obadia, N., Magnin, M. & Garcia-Larrea, L. On the importance of placebo timing in rTMS studies for pain relief. *Pain* **152**, 1233–1237 (2011).
98. Zunhammer, M. et al. The effects of treatment failure generalize across different routes of drug administration. *Sci. Transl. Med.* **9**, eaal2999 (2017).
99. Kessner, S., Wiech, K., Forkmann, K., Ploner, M. & Bingel, U. The effect of treatment history on therapeutic outcome: an experimental approach. *J. Am. Med. Assoc. Intern. Med.* **173**, 1468–1469 (2013).
100. Jenewein, J. et al. Fear-learning deficits in subjects with fibromyalgia syndrome? *Eur. J. Pain* **17**, 1374–1384 (2013).
101. Meulders, A. et al. Contingency learning deficits and generalization in chronic unilateral hand pain patients. *J. Pain.* **15**, 1046–1056 (2014).
102. Zaman, J., Vlaeyen, J. W., Van Oudenhove, L., Wiech, K. & Van Diest, I. Associative fear learning and perceptual discrimination: a perceptual pathway in the development of chronic pain. *Neurosci. Biobehav. Rev.* **51**, 118–125 (2015).
103. Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X. & Bishop, S. J. Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nat. Neurosci.* **18**, 590–596 (2015).
104. Koban, L. et al. Social anxiety is characterized by biased learning about performance and the self. *Emotion* **17**, 1144–1155 (2017).
105. Rutledge, R. B., Skandali, N., Dayan, P. & Dolan, R. J. A computational and neural model of momentary subjective well-being. *Proc. Natl Acad. Sci. USA* **111**, 12252–12257 (2014).
106. Eldar, E. & Niv, Y. Interaction between emotional state and learning underlies mood instability. *Nat. Commun.* **6**, 6149 (2015).
107. Jepma, M., Jones, M. & Wager, T. D. The dynamics of pain: evidence for simultaneous site-specific habituation and site-nonspecific sensitization in thermal pain. *J. Pain.* **15**, 734–746 (2014).
108. Wager, T. D. et al. Brain mediators of cardiovascular responses to social threat. Part II: Prefrontal-subcortical pathways and relationship with anxiety. *Neuroimage* **47**, 836–851 (2009).
109. Wager, T. D. et al. Brain mediators of cardiovascular responses to social threat. Part I: Reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. *Neuroimage* **47**, 821–835 (2009).
110. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
111. Zajkowski, W. K., Kossut, M. & Wilson, R. C.A. A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife* **6**, e27430 (2017).
112. Jones, M., Curran, T., Mozer, M. C. & Wilder, M. H. Sequential effects in response time reveal learning mechanisms and event representations. *Psychol. Rev.* **120**, 628–666 (2013).
113. Sutton, R. S. Gain adaptation beats least squares? In *Proc. 7th Yale Workshop on Adaptive and Learning Systems* 161–166 (1992); https://pdfs.semanticscholar.org/7ec8/876f219b3b3d5c894a3f395c89c382029cc5.pdf
114. Yu, A. & Cohen, J. in *Advances in Neural Information Processing Systems* Vol. 22 (eds Bengio, Y. et al.) 1873–1880 (NIPS Foundation, La Jolla, 2009).
115. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
116. Jacobs, R. A. Optimal integration of texture and motion cues to depth. *Vision Res.* **39**, 3621–3629 (1999).
117. Kakade, S. & Dayan, P. Acquisition and extinction in autoshaping. *Psychol. Rev.* **109**, 533–544 (2002).
118. Kording, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
119. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–29 (2017).
120. Gelman, A. *Bayesian Data Analysis* 3rd edn (CRC Press, Boca Raton, 2014).
121. Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 515–534 (2006).
122. Bennett, C. H. Efficient estimation of free-energy differences from monte-carlo data. *J. Comput. Phys.* **22**, 245–268 (1976).
123. Meng, X. L. & Wong, W. H. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sin.* **6**, 831–860 (1996).
124. Gronau, Q. F. et al. A tutorial on bridge sampling. *J. Math. Psychol.* **81**, 80–97 (2017).
125. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
126. Wager, T. D., Keller, M. C., Lacey, S. C. & Jonides, J. Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage* **26**, 99–113 (2005).

Acknowledgements

We thank M. Powell and D. Ryan for assistance with data collection, and M. Roy and M. López-Solà for discussions. This research was made possible with the support of National Institutes of Health grants NIMH 2R01MH076136 and R01DA027794 (to T.D.W.), a VENI grant of the Netherlands Organization for Scientific Research (to M. Jepma), and AFOSR grant FA9550-14-1-0318 (to M. Jones). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions

M.Jepma, L.K. and T.D.W. conceived and designed the experiments. M.Jepma conducted the experiments and analysed the data. L.K., J.D., M.Jones and T.D.W. provided expertise and feedback. M.Jepma, L.K., J.D., M.Jones and T.D.W. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0455-8.

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

Correspondence and requests for materials should be addressed to M.J.

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

nature research

Corresponding author(s): Marieke Jepma

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	E-Prime 2.0 (PST Inc.)
Data analysis	Matlab R2015b SPM8 custom fMRI analysis code, available at http://wagerlab.colorado.edu/tools

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The single-trial behavioral and NPS data are available through the Open Science Framework repository, https://osf.io/bqkz3/

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We report one behavioral and one fMRI experiment. Data are quantitative.
Research sample	Healthy volunteers, 18-55 years old, 25% female in Study 1 and 50% female in Study 2.
Sampling strategy	Convenience/availability sampling (participants signed up to participate, and were included if they fulfilled the inclusion criteria). We tested 30 (28 after exclusion of 2 participants, see below) and 34 participants in Study 1 and 2, respectively. We chose sample sizes of ~30 as these provide approximately 80% power to detect an effect size of Cohen's d = 0.54 or larger, an effect size near the lower bound of what we would expect based on previous heat-pain conditioning studies conducted in our lab.
Data collection	Behavioral (expectation and pain rating) data were collected via a computerized visual analogue scale, and fMRI data were collected using a Siemens 3T Trio scanner. No one else was present in the testing room besides the participant (and, in Study 1, the experimenter). The experimenter was not blind to the study hypothesis during data collection.
Timing	Study 1: 15 November 2013 - 31 January 2014. Study 2: 28 April 2014 - 8 December 2014
Data exclusions	In Study 1, one participant had to be excluded because of thermode failure. In Study 2, one participant misunderstood the expected-pain rating procedure, and was excluded from all analyses involving pain expectations.
Non-participation	In Study 1, one participant decided to stop before the end of the experiment because she found the heat too painful.
Randomization	Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants were recruited through fliers on university bulleting boards, university mailing lists, newspaper ads and online bulletin boards (such as craigslist).

Magnetic resonance imaging

Experimental design

Design type	task, event-related

Design specifications	70 test trials per subject, split up in 5 runs of 14 trials. Trials lasted 26-30 seconds, and were separated by inter-trial intervals of 3.5-7.5 seconds. Runs were separated by breaks of a few minutes during which we moved the thermode to a new spot on the subject's leg.
Behavioral performance measures	expectation and pain ratings on each trial were recorded.

Acquisition

Imaging type(s)	functional MRI
Field strength	3T
Sequence & imaging parameters	Echo-planar imaging sequence. TR = 1300 ms, TE = 25 ms, field of view = 220 mm, 3.4×3.4×3.0 mm voxels, 26 slices, parallel imaging, SENSE factor 2.
Area of acquisition	whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Functional images were slice-acquisition-timing and motion corrected using SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK). Structural T1-weighted images were coregistered to the first functional image for each subject using an iterative procedure of automated registration using mutual information coregistration in SPM8 and manual adjustment of the automated algorithm's starting point until the automated procedure provided satisfactory alignment. Structural images were normalized to MNI space using SPM8, interpolated to 2×2×2 mm voxels, and smoothed using a 6mm full-width at half maximum Gaussian kernel. We discarded the first 6 volumes of each run, and then concatenated the 5 test-phase runs for each participant. A high-pass filter of 180 seconds was used.
Normalization	Structural images were normalized to MNI space using SPM8, interpolated to 2×2×2 mm voxels, and smoothed using a 6mm full-width at half maximum Gaussian kernel
Normalization template	MNI152
Noise and artifact removal	Regressors of non-interest (nuisance variables) were i) "dummy" regressors coding for each run (intercept for each but the last run); ii) linear drift across time within each run; iii) the 6 estimated head movement parameters (x, y, z, roll, pitch, and yaw), their mean-zeroed squares, their derivatives, and squared derivatives for each run (total 24 columns per run); iv) indicator vectors for outlier time points identified based on their multivariate distance from the other images in the sample (see below); v) indicator vectors for the first two images in each run. Identification of outlier time points: Prior to preprocessing, global outlier time points (i.e. "spikes" in the BOLD signal) were identified by computing both the mean and the standard deviation (across voxels) of values for each image for all slices. Mahalanobis distances for the matrix of slice-wise mean and standard deviation values (concatenated) x functional volumes (time) were computed, and any values with a significant $\chi 2$ value (corrected for multiple comparisons based on the more stringent of either false discovery rate or Bonferroni methods) were considered outliers. The output of this procedure was used as a covariate of noninterest in the first-level models.
Volume censoring	none

Statistical modeling & inference

Model type and settings	We conducted first-level GLM analyses in SPM8, employing the single trial, or "single-epoch", design and analysis approach (e.g., Mumford, Davis, & Poldrack, 2014). To estimate single-trial pain responses, we constructed a GLM design matrix with separate regressors for the 1.8-s heat-application period in each trial (modeled as boxcar regressors, convolved with the canonical hemodynamic response function; other relevant trial events were modeled as well).
Effect(s) tested	We computed the expression of the neurological pain signature (NPS; a multivariate pattern of fMRI activity that was found to be sensitive and specific to physical pain in previous studies; Wager et al., 2013) on each trial, and tested for effects of cue type (high- vs. low-pain cue), heat intensity (temperature), the temperature x cue type interaction, the linear and quadratic effects of time (i.e., trial number) and their interactions with cue type, and the linear and quadratic effects of site-specific repetition.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	voxel-level
Correction	We used False discovery rate (FDR) to correct for multiple comparisons in whole-brain voxel-wise analyses.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

We applied a previously developed multivariate fMRI pattern (the NPS, see above) to the data in this study.