**Results**

**Computational models**

To formalize and further test the mechanisms underlying the observed cue effects on pain perception and learning, we applied two computational models: A computational process model and a Bayesian model (see Methods for equations). Both models contain perceptual inference and learning mechanisms. In addition, both models contain a confirmation bias on learning, implemented as a persistent influence of initial expectations on learning. The computational process model represents expectations and noxious input as point values, which impact on pain perception and learning is controlled by (constant) free parameters. The Bayesian model represents expectations and noxious input as probability distributions, which relative precisions (i.e., inverse variance) determine their trial-specific impact on perception and learning.

Model 1: Computational process model

This model has two levels (Figure 5). Level 1 is a perceptual inference mechanism which assumes that the pain evoked by a noxious stimulus is a weighted average of the current noxious stimulus intensity and cue-based expectation. The weighting of nociceptive input vs. expectation is controlled by parameter *γ*. Higher *γ* yields stronger impact of expectations on pain. Level 2 is a learning mechanism: The perceived pain triggers the updating of the relevant cue’s pain expectation according to a standard reinforcement-learning algorithm (delta rule)32. 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 experience-resistant expectations. To allow for a confirmation bias in learning, two learning rates *α*c and *α*i, control updating when the sign of prediction errors (pain - expected pain) is consistent or inconsistent, respectively, with the initial low or high pain association of the preceding cue. If *α*c is higher than *α*ithis implies a confirmation bias. Thus, this model has three free parameters: *γ*, *αc* and *αi*. Parameter *γ* controls the impact of the current expectation on pain, and *αc* and *αi* govern learning in a way that is potentially shaped by initial beliefs.

Model 2: Bayesian model

This model represents expectations, nociceptive input, and pain as Gaussian probability distributions (cf.16,54,55). Like Model 1, it has two levels. At Level 1, the current nociceptive input and cue-based expectation (prior) distributions jointly determine the perceived pain distribution. According to optimal Bayesian inference, the mean of the pain distribution is a precision-weighted average of the means of the nociceptive input and prior distributions, and the pain distribution is more precise (lower variance) than either the nociceptive input or the prior distribution. At Level 2, the perceived pain distribution triggers the updating of the relevant cue’s expectation distribution, again according to optimal Bayesian inference. Thus, the mean of the updated prior distribution is a precision-weighted average of the prior and perceived pain distributions on the previous trial, and the updated prior distribution is more precise than either the previous prior or the perceived pain distribution.

The iteration of these Bayesian inference rules across trials produces increasingly precise expectations, resulting in an increasingly strong impact of expectations on pain and increasingly less experience-driven expectation updating. Therefore, this model predicts that, after a few trials, pain is fully determined by expectations and expectation are no longer updated. To allow continuing effects of nociceptive input on pain and expectation updating, we made one adjustment to the model. Specifically, we allowed the variance of the perceived pain distribution at Level 1—presumably reflecting the uncertainty about whether the perceived pain is representative of future cue-pain associations—to change. The direction and amount of change are controlled by free parameter . If > 0 the pain distribution’s variance increases, resulting in weaker expectation updating and a less precise expectation at Level 2. If < 0 the pain distribution’s variance decreases, resulting in stronger expectation updating and a more precise expectation at Level 2. To allow for a confirmation bias in learning, we estimated separately for trials on which the sign of prediction errors is consistent and inconsistent with the initial low or high pain association of the preceding cue: c and i, respectively. If c is lower than i this implies a confirmation bias. Besides c and i, the ratio between the initial variance of the expectation distributions and the variance of the nociceptive input distribution (; see Methods) is a free model parameter.

**Methods**

**Computational models**

We developed two hierarchical computational models capturing perceptual inference, and (biased) learning. The models were implemented in Matlab (R2012b; Mathworks, Natick, MA). Model 1 represents noxious input, expectations, and perceived pain as discrete values, and can capture resistant expectations through a confirmation bias in learning. Model 2 represents noxious input, expectations, and perceived pain as Gaussian distributions—modeling perceptual inference and learning as Bayesian integration processes—and can capture resistant expectations by allowing expectations to decay to their original value after each learning episode.

*Model 1*

This model has two levels. Level 1 (Equation 1.1) is a perceptual inference mechanism that models the perceived pain on each trial *t* (Pt) as a combination of the current noxious input (Nt) and cue-based pain expectation (Ec,t), where the subscript [*c, t*] indicates expectations for cue *c* on trial *t*. Level 2 (Equations 1.2 and 1.3) is a learning mechanism that governs the updating of cue-based expectations in response to new pain outcomes. The original cue-based expectations can govern learning rate through a confirmation-bias mechanism (Equation 1.4).

At Level 1, the pain evoked by the noxious stimulus on trial *t* is a weighted linear combination of the intensity of that noxious stimulus and the pain expectation associated with the preceding cue:

Pt = (1-γ)Nt + γEc,t [1.1]

Parameter *γ* (0 ≤ *γ* ≤ 1) controls the relative contributions of Nt and Ec,t, such that the impact of expectations on perceived pain increases with increasing *γ*.

At Level 2, the discrepancy between Pt and Ec,t elicits a prediction error (PE):

PEt = Pt – Ec,t [1.2]

This prediction error triggers the updating of the pain expectation for cue *c*, according to a standard reinforcement-learning algorithm32:

Ec,t+1 = Ec,t + αPEt , which can also be written as Ec,t+1 = αPt  + (1-α)Ec,t [1.3]

The amount of updating is determined by learning rate parameter *α.* Higher values of *α* result in stronger updating.

We implemented a confirmation bias in learning rate by specifying *α* depending on whether the sign of a prediction error is consistent or inconsistent with the preceding cue’s initial low or high pain association:

[1.4]

chigh and clow refer to cues that were initially associated with high and low pain, respectively. Thus, *αc* and *αi* control expectation updating when the sign of a PE is, respectively, consistent and inconsistent with the cue’s initial high or low pain association. If αc is higher than αi, this implies a confirmation bias.

This model has three free parameters: *γ*, *αc* and *αi*. Although learning rates are usually constrained to the interval [0,1], the average learning rate directly computed from participants’ rating data was higher than 1 in some conditions in Study 2 (Figure 4A). Therefore, we constrained *αc* and *αi*to the interval [0,10].

*Model 2*

This model also has two levels, containing perceptual-inference (Equations 2.1-2.4) and learning (Equations 2.5-2.7) mechanisms, respectively. However, unlike Model 1, this model represents cue-based expectations (priors) and noxious input as Gaussian distributions *N*() and *N*(), respectively.

At level 1, the cue-specific prior and the noxious input distributions on trial *t* are combined into a perceived pain distribution, according to Bayes optimal weighting. The mean of the perceived pain distribution is:

+(1- [2.1]

controls the relative weights of the prior and noxious input distributions. The value ofis in turn determined by the relative variances of the prior and noxious-input distributions, such that the distribution with the lowest variance (i.e., highest precision) has the strongest impact on perceived pain:

/ ( [2.2]

The variance of the perceived pain distribution is:

() + [2.3]

Parameter allows the variance of the perceived pain distribution to change, effectively modulating the learning rate at Level 2 (see below, equation 2.5). We implemented a confirmation bias in learning rate by specifying depending on whether the sign of ( - ), i.e., the pain prediction error, is consistent or inconsistent with the preceding cue’s initial low or high pain association:

[2.4]

If c is lower than i this implies a confirmation bias

At level 2, the cue-based expectation is updated toward the perceived pain outcome, again according to Bayes optimal weighting, such that mean of the updated expectation is:

+(1- [2.5]

controls how much the expectation is updated in response to the perceived pain on trial *t*. Higher values of result in stronger updating. The value of is determined by the relative variances of the prior and perceived pain distributions:

/ ( [2.6]

The variance of the updated cue-based expectation is:

() [2.7]

Note that the posterior expectation is more precise (lower variance) than either the prior expectation or the perceived pain distributions.

This model had three free parameters: the ratio between the initial variance of the expectation distributions and the variance of the noxious input distribution,c, and i. We assumed that is constant over trials, and that all cues start out with equal prior variances . As the absolute scale of the variances is arbitrary, we set to a value of 1, and estimated .

*Model fitting*

We fitted both models to each participant’s trial-to-trial sequences of expectation and pain ratings using Matlab’s fmincon function, a constrained nonlinear optimization algorithm, with one thousand randomized parameter starting points. Specifically, for Model 1, we determined the parameter values that simultaneously minimized the sum of squared errors (SSE) between the model-generated values of Ec,t and the participant’s expected-pain ratings, and between the model-generated values of Pt and the participant’s pain ratings. Similarly, for Model 2, we determined the parameter values that simultaneously minimizing the SSE between the model-generated values of and the participant’s expected-pain ratings, and between the model-generated values ofand the participant’s pain ratings. To optimize fits, we initialized the means of the cue-based expectation distributions (Ec,1 and in Models 1 and 2, respectively) to the first expected-pain rating following each cue in the test phase, separately for each participant. In addition, we set the mean of the noxious-input distribution on each trial (Nt and in Models 1 and 2, respectively) to the average pain rating for the temperature presented on that trial.

We also examined the fit of three reduced versions of Model 1 (Table 1) that did not include expectancy-based pain modulation (*γ* = 0), did not include a confirmation bias (*α*c = *α*i), or did not include either of those two mechanisms. In addition, we examined the fit of a reduced version of Model 2 that did not include a confirmation bias (c = i).

****

**Figure 5**. Computational model capturing effects of cue-based expectations on pain and expectation updating. *γ* controls the relative impact of expectations and noxious input on perceived pain. *α*c and *α*i are the learning rates when the directions of prediction errors are, respectively, consistent and inconsistent with the preceding cue’s initial low or high pain association. The Bayesian model (Model 2) represents N, Ec and P as Gaussian distributions whose relative precisions (inverse variances) govern perceptual inference and expectation updating on each trial. It can capture a confirmation bias by increasing the variance of P (effectively reducing expectation updating) as a function of whether the direction of a prediction errors is consistent or inconsistent with the preceding cue’s low or high pain association.

Under the Kalman filter (KF) model, the subject uses Bayesian inference to infer the true amount of painful stimulation on each trial.

The KF model attributes to the subject a generative model of the experiment environment, under which the subject’s perceived noxious input on trial , denoted , is a Gaussian random variable centered on the true stimulus value () with observation variance :

(1)

The subject’s generative model also assumes random change in the stimulus value from one trial to the next, following a Gaussian random walk with variance :

(2)

Under the standard KF model, the subject performs optimal Bayesian inference with respect to (1) and (2), using a Gaussian conjugate prior. That is, the subject’s belief at the beginning of learning is given by

, (3)

where are technically free parameters but have little effect on model predictions except the first few trials. (For simplicity, they could be set to the true mean and variance of the stimulus sequence.)

Bayesian updating in the KF model follows a simple recursive equation. Denote the subject’s prior belief at the start of each trial as

(4)

where is the sequence of all inputs through trial . Then the posterior, after observing the input , is given by

(5)

Intuitively, the mean of the posterior is a weighted average of the two relevant sources of evidence, the mean of the prior belief () and the current input (), with each source weighted in proportion to its precision (i.e., inverse variance). The precision of the posterior is the sum of the precisions of the two sources.

The prior for the next trial is obtained from the previous posterior simply by adding the noise from the random walk:

(6)

Comparing (6) and (4) yields a recursive relationship that is the updating rule for the KF model:

(7a)

(7b)

Equations (3), (4), and (7) define the dynamics of the subject’s belief. What remains to complete the psychological model is a specification of the subject’s response.

The subject’s reported pain rating after experiencing the stimulus on trial is determined by the posterior belief, . One reasonable assumption is that the response will equal the mean of that posterior (). Alternatively, it might be assumed that the response is randomly sampled from the posterior (5).

Likewise, the subject’s expected pain prior to the stimulus on trial is determined by the prior belief, . Again, this response could be assumed to equal to mean of the distribution, , or it could be sampled from the full distribution, .

When there are multiple cues, the subject’s generative model is expanded to assume multiple random walks, , one for each cue . The subject maintains separate belief distributions for all cues, with means and variances . On trials when cue is presented (), the update rules match those in (7):

(8a)

(8b)

For the non-presented cue (), the only change to the subject’s belief comes from the random walk:

(9a)

(9b)

To model confirmation bias, we assume the subject deviates from Bayesian rationality by using a different value of depending on whether the update in (8a) moves toward or away from the relative associations learned in the training phase. That is, when the input disagrees with those associations, the subject discounts that input by treating is as being noisier.

More specifically, we assume the subject uses two values for the observation variance, > . The former value is used when and , and when and . The latter value is used when and , and when and .

Note that there are other ways for a KF model to predict enduring effects of the training phase, while remaining purely Bayesian. The most obvious is to put the bias into the random walk, as a spatial inhomogeneity. That is, we can include a directed component in the dynamics of the true stimulus value, replacing (2) with

(10)

where is the value associated to cue during the training phase. Thus the subject’s generative model assumes tends to decay toward . Consequently, the subject’s mean estimate exhibits a similar decay, with (8) replaced by

(11a)

(11b)

To give more intuition to the update rule for the mean, (11a) can be rewritten as

(12)with variable learning rate

. (13)

This learning rate is greater when is positive, meaning the current input is drawing in the direction of .