Knowing what you don't know: Estimating the uncertainty of feedforward and feedback inputs with prediction-error circuits

Loreen Hertäg[1,*], Katharina A. Wilmes[2], Claudia Clopath[3]

1 Modeling of Cognitive Processes, TU Berlin, Berlin, Germany.
2 Department of Physiology, University of Bern, Switzerland.
3 Bioengineering Department, Imperial College London, London, UK.

* loreen.hertaeg@tu-berlin.de

Abstract

At any moment, our brains receive a stream of feedforward sensory stimuli arising from the world we interact with. Simultaneously, neural circuits are shaped by feedback signals carrying predictions about the same feedforward inputs our brains are bombarded with. Those feedforward and feedback inputs often do not perfectly match. Thus, our brains have the challenging task of integrating these conflicting information according to their reliabilities. However, how neural circuits keep track of both the sensory and prediction uncertainties is not well understood. Here, we propose a network model whose core is hierarchical prediction-error circuits. We show that our network can estimate the variance of the sensory stimuli and the uncertainty in the prediction using the activity of negative and positive prediction-error neurons. In line with previous hypotheses, we demonstrate that neural circuits rely strongly on the feedback predictions at the beginning of a new stimulus, and if the environment is stable and the sensory cues are noisy. In addition, we investigate the mechanisms underlying a pathological weighting of feedforward and feedback inputs. In our network, the uncertainty estimation, and, hence, how much we rely on predictions, can be biased by perturbing the intricate interplay of different inhibitory interneurons. Finally, we demonstrate the link to biased perception and unravel how the different types of uncertainty contribute to the contraction bias.

Introduction

To survive in an ever-changing environment, animals must flexibly adapt their behavior based on previously encoded and novel information. This adaptation is reflected in the information processing of neural networks underlying context-dependent behavior. For instance, when you walk down an unknown staircase in a fully lit basement, your brain might entirely rely on the feedforward (bottom-up) input your senses receive (Fig. 1A, left). In contrast, when you walk down the same stairs in complete darkness, your brain might rely entirely on feedback (top-down) signals generated from a staircase model it has formed over previous experiences (Fig. 1A, middle). But how do neural networks switch between a feedforward-dominated and a feedback-dominated processing mode? And how do neural networks in the brain combine both input streams wisely? For instance, if you hike down an unexplored mountain in very foggy conditions, your brain receives unreliable visual information. In addition, it can only draw on a shaky prediction about what to expect (Fig. 1A, right).

A common hypothesis is that the brain weights different inputs according to their reliabilities. A prominent example of this hypothesis is Bayesian multisensory integration (see, e.g., Deneve and Pouget, 2004). According to this theory, neural networks represent information from multiple modalities by a linear combination of the uncertainty-weighted single-modality estimates. Multisensory integration is supported by several observations showing that animals can combine information from different modalities in a fashion that minimizes the variance of the final estimate (Ernst and Banks, 2002; Battaglia et al., 2003; Körding and Wolpert, 2004; Alais and Burr, 2004; Rowland et al., 2007; Gu et al., 2008; Fetsch et al., 2012). Here, we propose that the same concepts could be employed for the weighting of sensory inputs and predictions thereof (Körding and Wolpert, 2004; Yon and Frith, 2021). A central point in the weighting of inputs is the estimation of their variances as a measure of uncertainty. However, how the variance of both the sensory input and the prediction can be computed on the circuit level is not resolved yet.

We hypothesized that prediction error (PE) neurons provide the basis for the neural computation of variances. PEs are an integral part of the theory of predictive processing which states that the brain constantly compares incoming sensory information with predictions. If those predictions are wrong, the resulting PEs allow the network to revise the model of the world, thereby ensuring that the predictions become more accurate (Keller and Mrsic-Flogel, 2018). Experimental evidence suggests that these PEs

may be represented in the activity of distinct groups of neurons, termed PE neurons (Eliades and Wang, 2008; Keller and Hahnloser, 2009; Ayaz et al., 2019; Audette et al., 2021). Moreover, these neurons may come in two types when excitatory neurons exhibit near-zero, spontaneous firing rates (Rao and Ballard, 1999; Keller and Mrsic-Flogel, 2018). Negative PE (nPE) neurons only increase their activity when the prediction is *stronger* than the sensory input, while positive PE (pPE) neurons only increase their activity when the prediction is *weaker* than the sensory input. Indeed, it has been shown that excitatory neurons in rodent primary sensory areas can encode negative or positive PEs (Keller et al., 2012; Attinger et al., 2017; Jordan and Keller, 2020; Audette et al., 2021).

Here, we show that the unique response patterns of nPE and pPE neurons may provide the backbone for computing both the mean and the variance of sensory stimuli. Furthermore, we suggest a network model with a hierarchy of PE circuits to estimate the variance of the prediction, in addition to the variance of the sensory inputs. We show that in line with the ideas of multisensory integration, predictions are weighted more strongly than the sensory stimuli when the environment is stable (that is, predictable) and the sensory inputs are noisy. Moreover, we find that predictions are taken into account more at the beginning of a new trial than at the end, especially when the new sensory stimulus is reliable. In addition, we unravel the mechanisms underlying a neuromodulator-induced shift in the weighting of sensory inputs and predictions. In our model, these neuromodulators activate groups of inhibitory neurons such as parvalbumin-expressing (PV), somatostatin-expressing (SOM), and vasoactive intestinal peptide-expressing (VIP) interneurons (Markram et al., 2004; Rudy et al., 2011; Pfeffer et al., 2013; Jiang et al., 2015; Tremblay et al., 2016; Campagnola et al., 2022). These interneurons have been suggested to establish a multi-pathway balance of excitation and inhibition that is the basis for nPE and pPE neurons (Hertäg and Sprekeler, 2020; Hertäg and Clopath, 2022). By breaking this balance, the PE neurons change their baseline firing rate and gain, leading to a biased variance estimation. Finally, we show that this weighting can be understood as a neural manifestation of the contraction bias (Hollingworth, 1910; Jazayeri and Shadlen, 2010; Ashourian and Loewenstein, 2011; Petzschner and Glasauer, 2011; Akrami et al., 2018; Meirhaeghe et al.).

Results

Prediction-error neurons as the basis for estimating the mean and variance of sensory stimuli

We hypothesize that the distinct response patterns of negative and positive prediction-error (nPE/pPE) neurons act as a backbone for estimating the mean and the variance of sensory stimuli. An nPE neuron only increases its activity relative to a baseline when the sensory input is weaker than predicted, while a pPE neuron only increases its activity relative to a baseline when the sensory input is stronger than predicted. Moreover, both nPE and pPE neurons remain at their baseline activities when the sensory input is fully predicted (Fig. 1B). If the prediction equals the mean of the sensory stimulus, the PE neurons, hence, encode the deviation from the mean. Thus, the squared sum of nPE and pPE neuron activity represents the variance of the feedforward input (provided that the PE neurons are silent without sensory stimulation).

To test our hypothesis, we study a rate-based mean-field network. The core network is a prediction-error (PE) circuit with excitatory nPE and pPE neurons, as well as inhibitory parvalbumin-expressing (PV), somatostatin-expressing (SOM), and vasoactive intestinal peptide-expressing (VIP) interneurons (Fig. 1B). While the excitatory neurons are simulated as two coupled point compartments to emulate the soma and dendrites of elongated pyramidal cells, respectively, all inhibitory cell types were modeled as point neurons. The connectivity of and inputs to the network were chosen such that the excitatory (E) and inhibitory (I) pathways onto the pyramidal cells were partially balanced. This balance that is only temporarily broken during mismatches has been shown to be necessary for nPE and pPE neurons to emerge (Hertäg and Sprekeler, 2020; Hertäg and Clopath, 2022, see Methods).

We assume that this core circuit is shaped by feedback connections (Larkum, 2013; Harris and Shepherd, 2015) that have been hypothesized to carry information about expectations or predictions (Mumford, 1992; Larkum, 2013; Friston, 2008). To account for predictions, we model a memory (M) neuron that integrates the activity of the PE neurons (Fig. 1C). Following Keller and Mrsic-Flogel (2018), we assume that the pPE neuron excites the memory neuron, while the nPE neuron inhibits this neuron (for instance, through lateral inhibition, here not modeled explicitly). Because feedback connections are shown to target the apical dendrites of pyramidal cells (Larkum, 2013) and interneurons located in superficial layers of the cortex (see, e.g. Tremblay et al., 2016), the M neuron makes connections with the

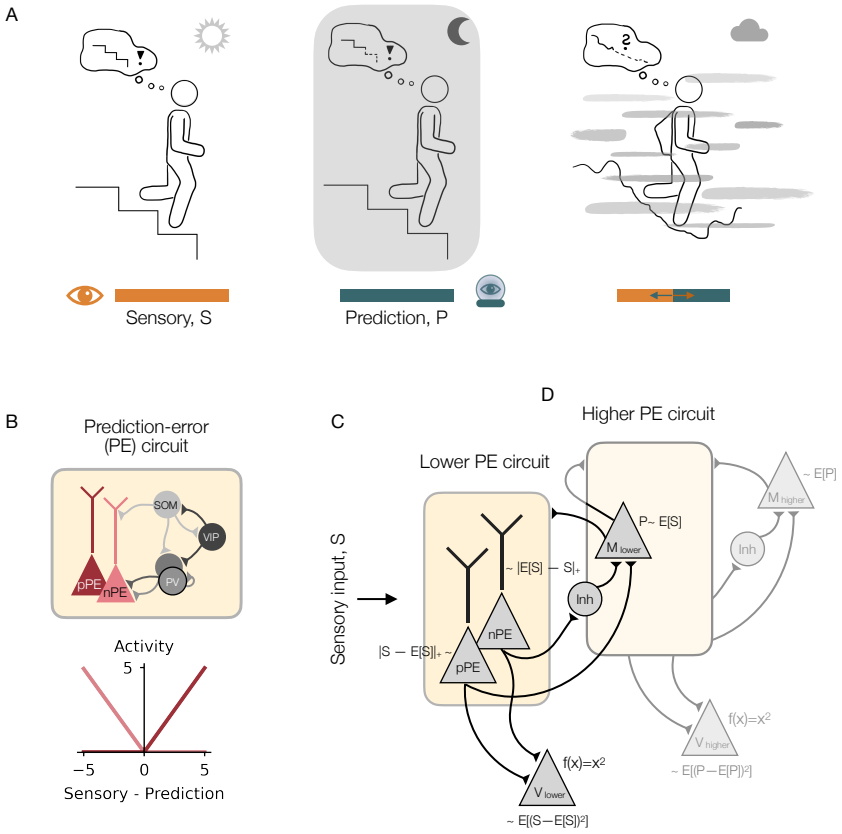

Figure 1. Neural network model to track both the uncertainty of sensory inputs and predictions. **(A)** Example illustration for context-dependent integration of information. Left: When walking down an unfamiliar staircase that is visible, the brain might rely solely on external sensory information. Middle: When walking down the same stairs without visual information, the brain might rely on predictions formed by previous experience. Right: When climbing down an unexplored mountain in foggy conditions, the brain might need to integrate sensory information and predictions simultaneously. **(B)** Illustration of a prediction-error (PE) circuit with both negative and positive PE (nPE/pPE) neurons that receive inhibition from three different inhibitory interneuron types: parvalbumin-expressing (PV), somatostatin-expressing (SOM), and vasoactive intestinal peptide-expressing (VIP) interneurons. Local excitatory connections are not shown for clarity. **(C)** Illustration of network model that estimates the mean and variance of the external sensory stimuli. The core of this network model is the PE circuit shown in (B). The lower-level V neuron encodes the variance, while the lower-level M neuron encodes the mean of the sensory input. **(D)** Same as in (C) but the feedforward input is the activity of the lower-level M neuron.

dendritic compartment of the PE neurons and some of the interneurons (here, VIP and PV neurons, see Methods for more details). We, furthermore, simulate a downstream neuron (termed V neuron), modeled as a leaky integrator with a quadratic activation function, that receives excitatory synapses from the PE neurons. Hence, in this setting, the V neuron encodes the variance of the sensory stimuli (Fig. 1C).

To show that this network can indeed represent the mean and the variance in the respective neurons, we stimulate it with a sequence of step-wise constant inputs drawn from a uniform distribution (Fig. 2A). We, hence, assume that the sensory stimulus varies over time. In line with the distinct response patterns for nPE and pPE neurons, these neurons change only slightly with increasing stimulus mean but increase strongly with input variance (Fig. 2B). In contrast, the three interneurons strongly increase with stimulus mean and only moderately increase with stimulus variance (Fig. 2C). The activity of the M neuron gradually approaches the mean of the sensory inputs (Fig. 2D, middle), while the activity of the V neuron approaches the variance of those inputs (Fig. 2E, middle). We show that this holds for a wide range of input statistics (Fig. 2D-E, right) and input distributions (Fig. S1). Small deviations from the true mean occur mainly for large input variances, while the estimated variance is fairly independent of the input statistics tested.

We verified our results in a heterogeneous network in which a population of neurons represents each neuron type of the PE circuit, and the synaptic connection strengths from each PE neuron onto the M and V neuron are different (see Methods, Fig. S2A). As before, the network can correctly estimate the mean and the variance of the sensory stimuli (Fig. S2B). Furthermore, we show that the errors with which the M and V neurons encode the stimulus statistics are independent of uncorrelated modulations of

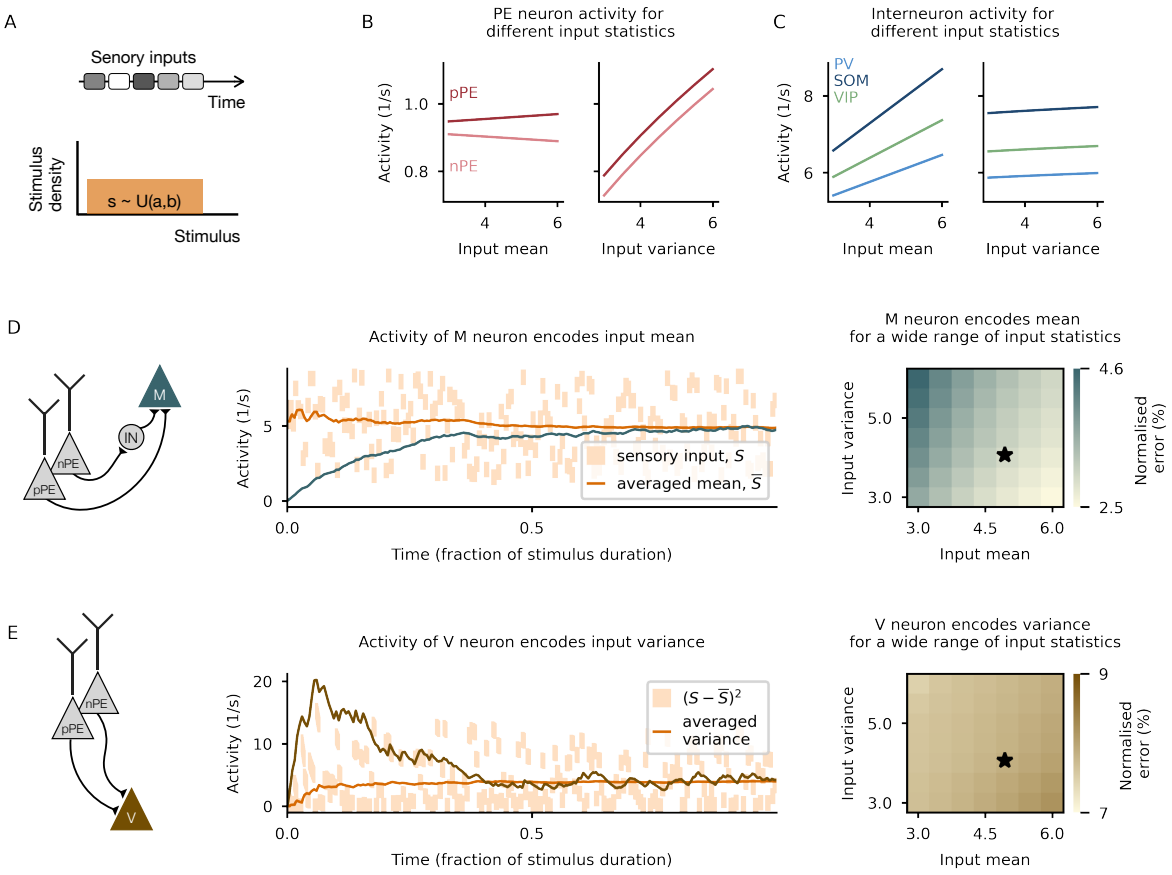

Figure 2. Prediction-error neurons as the basis for estimating mean and variance of sensory stimuli. **(A)** Illustration of the inputs with which the network (Fig. 1C) is stimulated. Network is exposed to a sequence of constant stimuli drawn from a uniform distribution. **(B)** PE neuron activity hardly changes with stimulus strength (left) but strongly increases with stimulus variability (right). **(C)** Interneuron activity strongly changes with stimulus strength (left) but hardly changes with stimulus variability (right). **(D)** M neuron correctly encodes the mean of the sensory stimuli. Left: Illustration of the input synapses onto the M neuron. Middle: Activity of the M neuron over time for one example distribution (black start in right panel). Right: Normalised absolute difference between the averaged mean and the activity of the M neuron in the steady state for different parametrizations of the stimulus distribution. **(E)** V neuron correctly encodes the variance of the sensory stimuli. Left: Illustration of the input synapses onto the V neuron. Middle: Activity of the V neuron over time for one example distribution (black start in right panel). Right: Normalised absolute difference between the averaged variance and the activity of the V neuron in the steady state for different parametrizations of the stimulus distribution.

those connection strengths (Fig. S2C) and the sparsity of the network (Fig. S2E). When all connection strengths are collectively shifted to higher values, the error increases for the V neuron, while it remains unaffected for the M neuron.

While our mean-field network was designed to track the mean and the variance of stimuli that vary in time, we reasoned that the same principles apply to stimuli that vary across space. To show that, we simulated a population network that consists of unconnected replicates of the mean-field network described above (Fig. S3A). Each mean-field network receives a short, constant input from a different part of the receptive field. If the connection strengths from the PE neurons to the M and V neurons are adjusted accordingly (see Methods), the network correctly estimates the stimulus average and spatial uncertainty. (Fig. S3B-C).

In summary, nPE and pPE neurons can serve as a basis to estimate the mean and the variance of sensory stimuli which vary over time and space.

Estimating the uncertainty of both the sensory input and the prediction requires a hierarchy of PE circuits

Following the ideas of Bayesian multisensory integration, the weighting of sensory stimuli and predictions would require knowledge about their uncertainties. As we have shown in the previous section, the variance of the sensory stimulus can be estimated using PE neurons. We hypothesize that the same principles

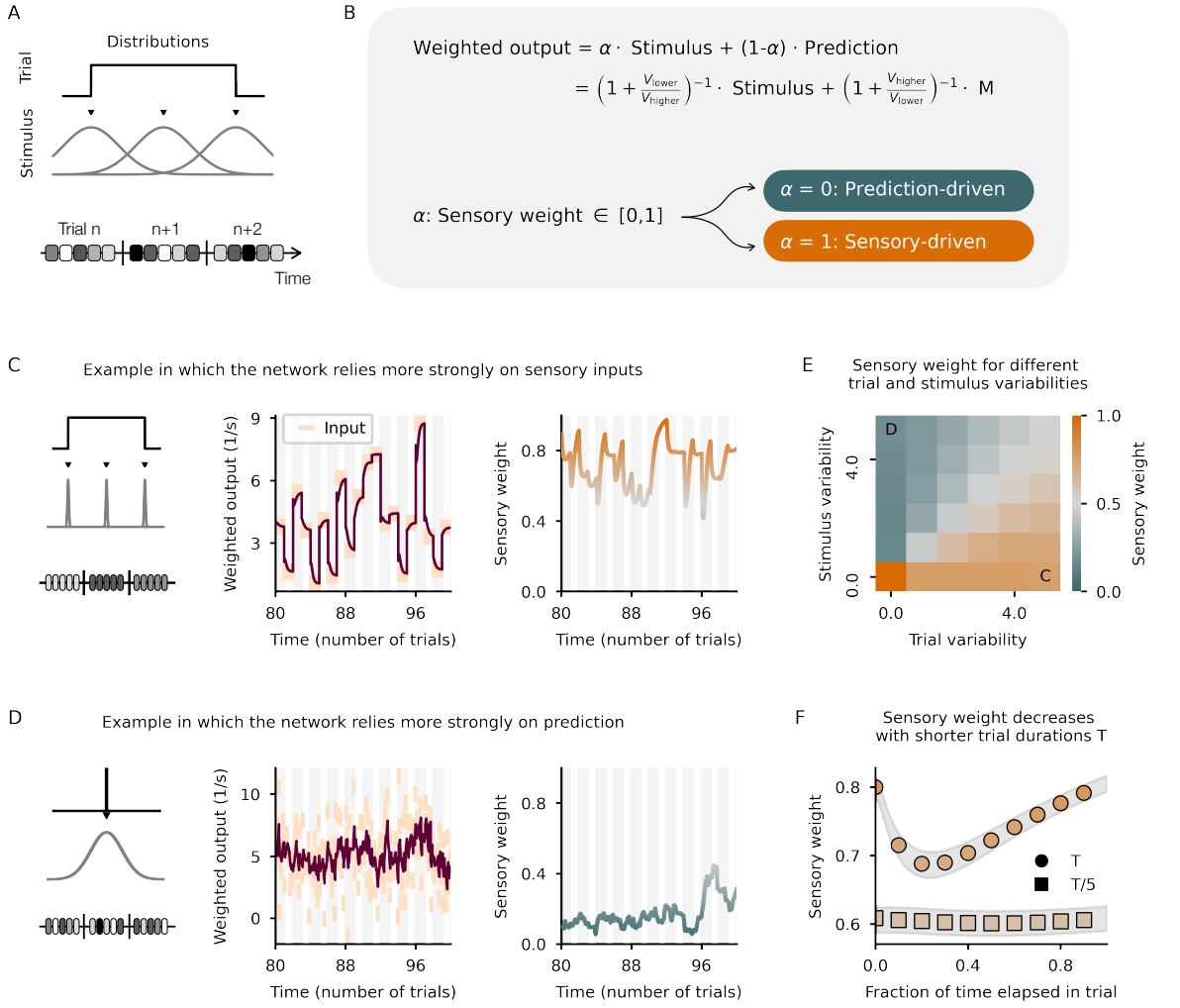

Figure 3. Estimating the uncertainty of both the sensory input and the prediction.
(A) Illustration of the stimulation protocol. The network is exposed to a sequence of stimuli (one stimulus per trial). To account for stimulus variability, each stimulus is represented by 10 stimulus values drawn from a normal distribution. To account for the volatility of the environment, in each trial, the stimulus mean is drawn from a uniform distribution (denoted trial variability). **(B)** Illustration of how the weighted output is calculated. The sensory weight α lies between zero (system relies perfectly on prediction) and one (system relies solely on the sensory input). **(C)** Limit case example in which the stimulus variability is zero but the trial variability is high. Left: Illustration of the stimulation protocol. Middle: Weighted output follows closely the sensory stimuli. Right: Sensory weight (function of the variances, see B) close to 1, indicating that the network ignores the prediction. Input statistics shown in E. **(D)** Limit case example in which the stimulus variability is high but the trial variability is zero. Left: Illustration of the stimulation protocol. Middle: Weighted output pushed towards the mean of the sensory stimuli. Right: Sensory weight close to zero, indicating that the network ignores the sensory stimuli. Input statistics shown in E. **(E)** Sensory weight for different input statistics. Predictions are weighted more strongly when the stimulus variability is larger than the trial variability. **(F)** Sensory weight throughout a trial for two different trial durations. Predictions are weighted more strongly at the beginning of a new trial.

apply to computing the variance of the prediction. Hence, we augment the network with a *higher* PE circuit that receives feedforward synapses from the M neuron of the *lower* PE circuit (Fig. 1D). Both subnetworks are identical except for the M neuron in the higher PE circuit which is modeled with slower dynamics than the one in the lower PE circuit.

To test the network's ability to estimate the variances correctly, we stimulated the network with a sequence of inputs whose mean can vary from trial to trial. More precisely, in each trial, the network is presented with a stimulus that is composed of n constant, consecutive values drawn from a normal distribution. The variance of this distribution represents the stimulus noise. To account for potential changes in the environment, we draw the stimulus mean from a uniform distribution (Fig. 3A). Hence, the inputs change on two different time scales, where the stimulus variability has a faster time scale than the trial variability.

Following the formalism of multisensory integration (see, e.g. Pouget et al., 2013), we assume that the

network's output is a weighted sum of the feedforward sensory input and the feedback prediction. The weights assigned to each input stream are functions of the uncertainties, that is, the activities of the V neurons. The sensory weight captures, hence, how much the network relies on the sensory input (Fig. 2B). To test our network, we first consider two limit cases. In the first limit case, a different, low-variance stimulus is presented in each trial (Fig. 3C, left). According to the theory, the network should follow the sensory inputs closely and ignore the predictions. When we arithmetically calculate the weighted output (Fig. 3C, middle) and the sensory weight (Fig. 3C, right), the network indeed shows a clear preference for the sensory input. In the second limit case, the same, high-variance stimulus is presented in each trial (Fig. 3D, left). According to the theory, the network should downscale the sensory feedforward input and weight the prediction more strongly. Indeed, the weighted output of the network shows a clear tendency to the mean of the stimuli (Fig. 3D, middle), also reflected in the low sensory weight (Fig. 3D, right).

To validate the network responses fully, we systematically varied the trial and stimulus variability independently. If both variances are similar, the sensory weight approaches *0.5*, reflecting equal contribution of the sensory input and the prediction to the weighted output. Only if both variances are zero, the network represents the sensory input perfectly. In line with the limit case examples above, if the stimulus variance is larger than the trial variance, the network weights the prediction more strongly than the sensory input (Fig. 3E). Because the network dynamically estimates the sensory and prediction uncertainty, the sensory weight changes when the input statistics shifts (Fig. S4).

Inspecting closely the dynamics of our network, we noticed that the prediction is typically weighted higher at the beginning of a new trial than in the steady state. This is particularly pronounced in a sensory-driven input regime (see Fig. 3C). This is further confirmed in simulations in which the trail duration was shortened (Fig. 3F). Our model makes therefore the following experimentally testable prediction: Sensory predictions influence neural activity more significantly in experiments that rely on fast stimulus changes.

It has been hypothesized, that some symptoms in psychiatric disorders like autism and schizophrenia can be ascribed to a pathological weighting of sensory inputs and predictions (Yon and Frith, 2021). We thus wondered which network properties might bias the estimation of the variances, and, consequently, the weighting of different input streams. We identified the time scales at which the M neurons incorporate new information as a decisive factor in the integration of inputs. To show this, we varied the weights from the PE neurons onto the lower-level M neuron. If the weights are too small (M updates too slowly), the system relies too much on feedback predictions. In contrast, if the weights are too large (M updates too fast), the system relies too much on the feedforward sensory information (Fig. S5A). While the speeds at which the activity of the M neurons evolve influence the weighting of inputs, the precise activation function of the V neurons is less pivotal. When we replaced the quadratic activation function with a linear, rectified function, the V neurons did not encode the variance but the average absolute deviation of the sensory stimuli. However, the sensory weight is only slightly shifted to larger values for low trial/high stimulus variability (Fig. S5B).

In summary, we show that the variances of both the sensory inputs and predictions thereof can be dynamically computed in networks comprising a lower and higher PE circuit. In such a network, predictions are given more weight at the beginning of a new stimulus, and if the sensory inputs are noisy while the environment is stable.

Biasing the weighting of sensory inputs and predictions by neuromodulators

The brain's flexibility and adaptability are supported by a plethora of neuromodulators which influence the activity of neurons in a variety of ways (Avery and Krichmar, 2017). A prominent target of neuromodulatory inputs is inhibitory neurons (Cardin, 2019; Hattori et al., 2017; Swanson and Maffei, 2019). Moreover, distinct interneuron types are differently (in-)activated by those neuromodulators (Wester and McBain, 2014; Hattori et al., 2017; Swanson and Maffei, 2019). We, therefore, wondered if and how the weighting of sensory inputs and predictions thereof may be biased when neuromodulators activate distinct interneuron types.

To this end, we modeled the presence of a neuromodulator by injecting an additional excitatory input into an interneuron type. We reasoned that the network effect of a neuromodulator not only depends on the interneuron type it targets but also on the inputs this neuron receives and the connections it makes with other neurons in the network. We, therefore, tested different mean-field networks that differ in the distribution of sensory inputs and predictions onto the interneurons, and the underlying connectivity. The commonality across those networks is that they exhibit an E/I balance of excitatory and inhibitory pathways onto the PE neurons (Hertäg and Clopath, 2022).

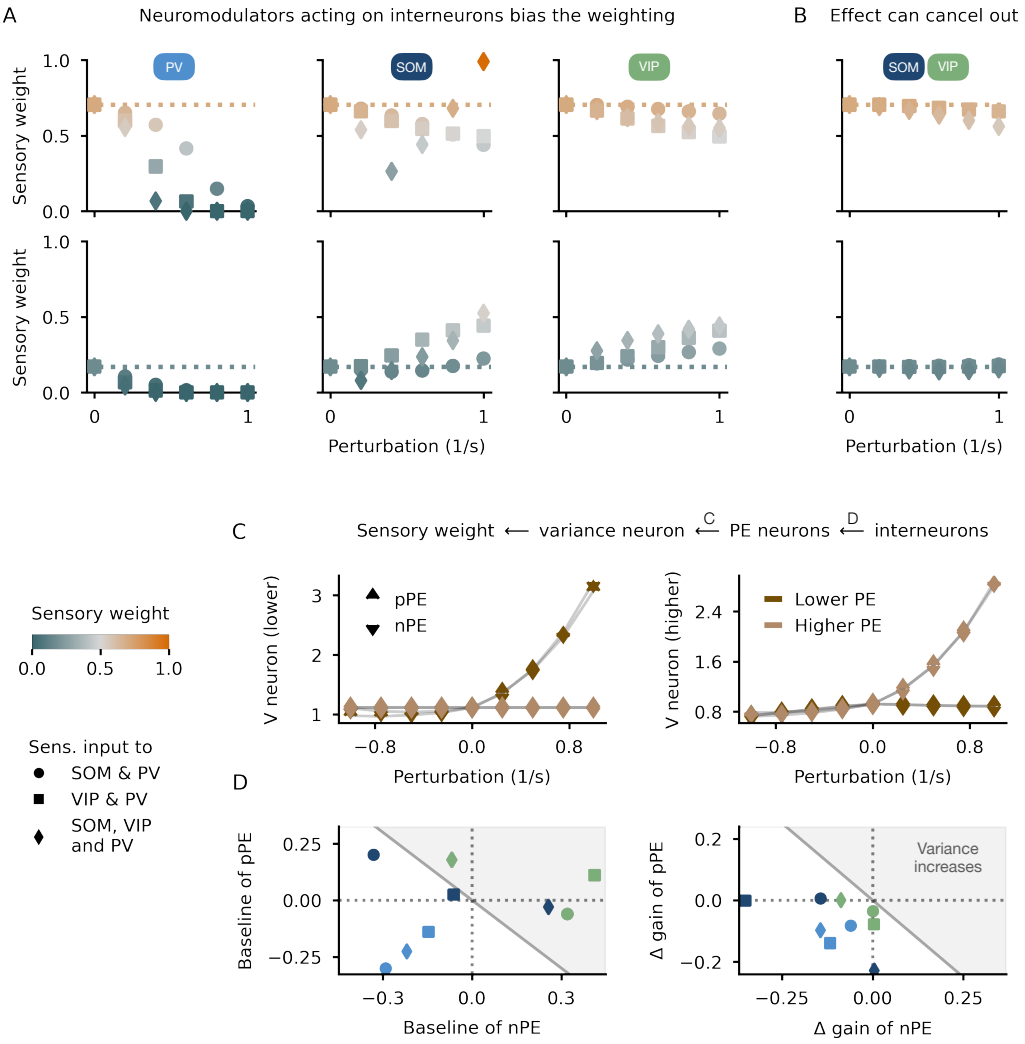

Figure 4. Neuromodulator-based shifts in the weighting of sensory inputs and predictions. **(A)** Neuromodulators acting on the interneurons can shift the weighting of sensory inputs and predictions. The changes depend on the type of interneuron targeted and the modulation strength (here simulated through an additional excitatory input). Considered are two limit cases (upper row: more sensory-driven before modulation, lower row: more prediction-driven before modulation). The results are shown for three different PE circuits (denotes by different markers). **(B)** When SOM and VIP neurons are equally modulated, the sensory weight remains unaffected. **(C)** The V neurons' activities depend on the PE neurons. Hence, perturbing the nPE and pPE neurons changes the uncertainty estimation. While stimulating the lower PE neurons affects both the lower and higher-order V neurons (right), stimulating the higher-order PE neurons only affects the V neuron in the same subnetwork (left). **(D)** The V neuron activity, and hence the sensory weight, changes as a result of the modulated PE neuron activity. The PE neuron activity, on the other hand, changes as a result of the interneurons being modulated. The interneurons change the baseline (left) and the gain (right) of the PE neurons. Whether an interneuron increases or decreases the estimated variance depends on both factors.

Across the different mean-field networks tested, increasing the activity of PV neurons biases the network's output toward predictions (Fig. 4A left). In contrast, increasing VIP activity forces the networks to weigh both inputs more equally. As a consequence, predictions are overrated in a sensory-driven input regime, and, sensory inputs are overrated in a prediction-driven input regime (Fig. 4A right). Increasing SOM neuron activity, while qualitatively similar to increasing VIP neuron activity, depends on the mean-field network tested and the strength of activation (Fig. 4A middle).

Neuromodulators are most likely increasing the activity of more than one interneuron type. To account for the co-activation of interneurons, we injected an excitatory input into two interneuron types at the same time and varied the strength with which each interneuron was modulated (Fig. S6). If PV neurons are the major target of a neuromodulator, the network is still biased toward predictions. If SOM and VIP neurons are equally stimulated, the weighting of sensory inputs and predictions remains largely unaffected (Fig. 4B), suggesting that the individual effects cancel out.

What are the network mechanisms underlying these observation? The sensory weight is a function of the lower and higher V neuron activity. Hence, any changes to the sensory weight result from changes to the neurons encoding the variances. In our network, the V neurons only receive excitatory synapses from PE neurons. Hence, any changes in the sensory weights upon activation of interneurons must be due to changes in the PE neurons. To disentangle the effect of nPE and pPE neurons, we perturbed those neurons individually in both the lower or higher subnetwork by injecting either an inhibitory or excitatory additional input (Fig. 4C). Stimulating either PE neuron in the lower subnetwork increases the activity of the lower-level V neuron strongly. Moreover, the higher-level V neuron is also slightly affected. This is because the lower-level M neuron is also modulated by the lower-level PE neurons and makes feedforward connections to the higher-level PE circuit. In contrast, stimulating either PE neuron in the higher subnetwork increases the activity of the higher-level V neuron but leaves the lower-level neurons unaffected.

This suggests that to understand the effect of neuromodulators on the sensory weight, we need to unravel the effect of interneuron activation on PE neurons. Increasing interneuron activity leads to changes in the baseline and gain of PE neurons that bias the estimation of mean and variance (Fig. S8, see Methods). In all three networks tested, activating PV neurons decreases both baseline and gain of the PE neurons, leading to a decrease in the estimated variance (Fig. 4D & Fig. S9). Stimulating the SOM or VIP neuron decreases the gain in either nPE or pPE neuron. However, the baseline of those neurons can either decrease or increase depending on the connectivity with other neurons in the network. The summed effect over nPE and pPE neuron (Fig. S9), hence, suggests that whether the activity of the V neuron increases or decreases depends on the input statistics: for low-mean stimuli, the elevated baseline activity dominates the changes in the variance, while for high-mean stimuli the changes in the gain dominate.

Altogether, we show that neuromodulators increasing the activity of interneurons bias the weighting of sensory inputs and predictions by changing the gain and baseline of PE neurons. Whether the sensory weight increases or decreases depends not only on the interneuron it targets but also on the network it is embedded in and the input regime.

Explaining the contraction bias with the weighting of sensory inputs and predictions

We hypothesized that the weighted integration of sensory inputs and predictions thereof manifests in all-day behavior, in the form of a phenomenon called *contraction bias*. The contraction bias describes the tendency to overestimate sensory stimuli drawn from the lower end of a stimulus distribution and to underestimate stimuli drawn from the upper end of the same distribution. This *bias toward the mean* has been reported in different species and modalities (Hollingworth, 1910; Jazayeri and Shadlen, 2010; Ashourian and Loewenstein, 2011; Petzschner and Glasauer, 2011; Akrami et al., 2018; Meirhaeghe et al.).

We first investigated whether the network's output can be interpreted as a neuronal manifestation of the contraction bias (see Methods for an illustrative analysis). To this end, we define the contraction bias as the trial-averaged difference between the weighted output and the sensory stimulus. When plotted over the trail-averaged stimuli, the bias is positive for stimuli below the mean of the input distribution and negative for stimuli above the mean (Fig. 5A), in line with a *bias toward the mean*. To measure the amount of bias in the network, we use the slope of the linear fit to the relationship between bias and trial stimulus. The larger the absolute slope, the larger the bias.

What are the underlying network factors that contribute to the neuronal contraction bias in the network? We have seen that how much the prediction is taken into account is determined by both the lower and higher-level V neurons encoding the variance of the stimulus and the prediction. Hence, the bias must be similarly influenced by these factors. When we increase the stimulus uncertainty, the bias increases (Fig. 5B). In contrast, when we increase the trial-to-trial uncertainty, the bias decreases (Fig. 5B).

To further disentangle the different sources of the bias, we first simulated a network without stimulus uncertainty (variance set to zero) for two trial-to-trail variances (volatility of the environment). In this case, the emerging contraction bias is independent of the volatility of the environment (Fig. 5C). We show mathematically that the bias results from the network output not yet reaching its new steady state within the trial duration (see Methods). In other words, the bias is the difference between the weighted output at the end of the trial and its steady state (the shown stimulus). How fast the new steady state is reached depends only on the time constants in the network and not the trial-to-trial variability. We

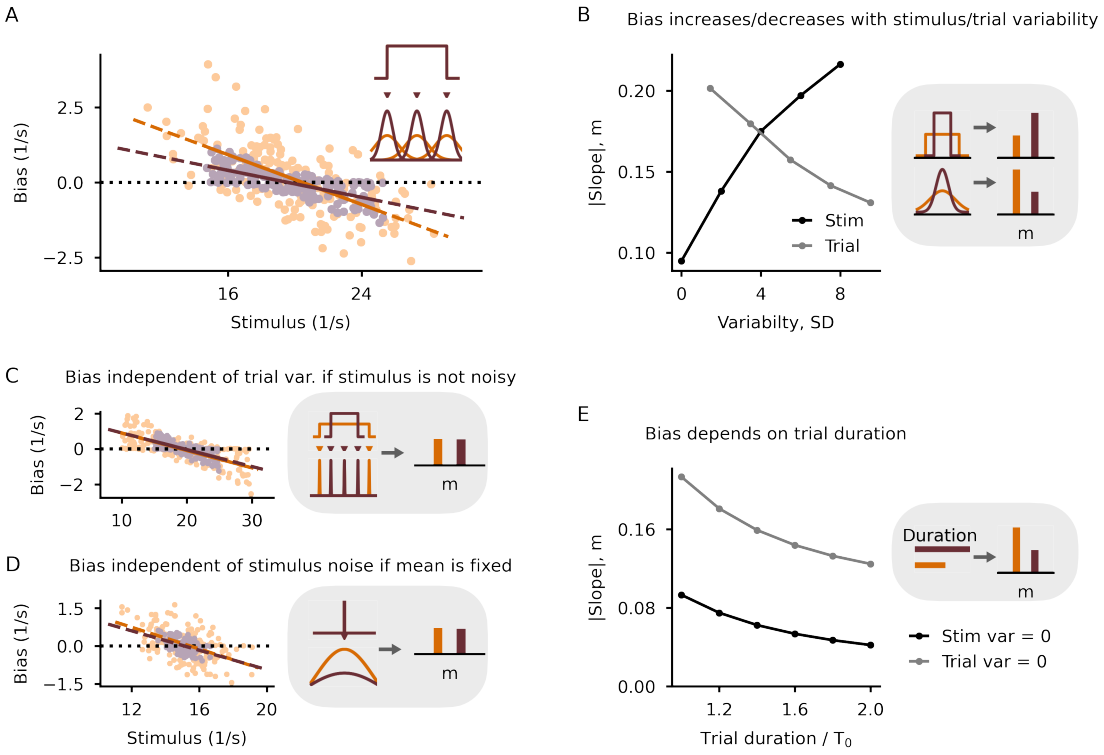

Figure 5. Mechanisms underlying the contraction bias.
(A) Contraction bias in the model for two different stimulus uncertainties depicted in the inset. Bias is defined as the weighted output minus the stimulus mean. The absolute value of the slope (see linear fit) is a measure of the bias. The larger the slope, the larger the bias. **(B)** As a consequence of the sensory weight, the slope increases with stimulus variability (bias increases) and decreases with trial variability (bias decreases). **(C)** Bias is independent of the trial variability when the stimulus variability is zero. **(D)** Bias is independent of the stimulus variability when the trial variability is zero. **(E)** The slope depends on the trial duration.

next resume the limit case in which the same, high-variance stimulus is shown in every trial. In this case, the contraction bias is also largely independent of the stimulus variance (Fig. 5D). Our mathematical analysis reveals that the bias is well described by the difference between the prediction, that is, mean stimulus over the history of all stimuli shown, and the current stimulus, weighted by a function of the trial duration. The analysis of both limit cases suggests that the bias also depends on the trial duration. To confirm this, we extended the trial duration for either limit case. As expected from the analysis, the bias decreases steadily in the simulations (Fig. 5E). Hence, we predict that the contraction bias can be reduced for sufficiently long trials.

So far, we assumed that the stimulus variance is independent of the stimulus mean. A consequence of this choice is that the bias on either end of the input distribution is largely the same (but with reversed signs). However, behavioral data (see, e.g. Rakitin et al., 1998) shows that the bias increases for stimuli drawn from the upper end of the distribution, a phenomenon usually attributed to *scalar variability*. To capture this in the model, we assume that the stimulus standard deviation linearly increases with the stimulus mean. In these simulations, as expected, the bias increases for a stimulus distribution shifted to higher trial means (Fig. S10).

In summary, we show that the weighted integration of sensory inputs and predictions can be interpreted as a neural manifestation of the contraction bias. While the stimulus and trial-to-trial variability shape the contraction bias, their contributions differ. Moreover, we reveal that the trial duration contributes to the bias.

Discussion

Our work has been driven by the puzzling question of how the brain integrates top-down feedback predictions with the sensory feedforward bottom-up inputs it constantly receives during behavior. This task may be particularly challenging when the prediction and the sensory information differ (Han and Helmchen, 2023). Conflicting information may be caused by noise in the sensory feedforward inputs or by

changes in the environment that could not be predicted. A prominent hypothesis is that how much we rely on our predictions and new sensory evidence is determined by an intricate balance between both, based on how reliable they are (see e.g. Körding and Wolpert, 2004; Yon and Frith, 2021).

This idea is consistent with Bayesian theories on the optimal integration of multiple sensory cues (aka multisensory integration). Ernst and Banks (2002) showed that to estimate the height of a bar humans combine visual and haptic information in a fashion that minimizes the variance of the final estimate. Similar studies confirmed that animals can optimally combine multiple sensory information by taking into account their uncertainties (Battaglia et al., 2003; Körding and Wolpert, 2004; Alais and Burr, 2004; Rowland et al., 2007; Gu et al., 2008; Fetsch et al., 2012). These behavioral studies were accompanied by neural recordings identifying populations of neurons that can form the basis of multisensory integration (Wallace et al., 1998; Gu et al., 2008; Fetsch et al., 2012).

Summary of findings

While multisensory integration is concerned with the weighting of sensory cues coming from different modalities, it is conceivable that the same ideas hold for top-down predictions and bottom-up sensory inputs that are combined in cortical circuits (Körding and Wolpert, 2004; Yon and Frith, 2021). Indeed, there is evidence that the brain can rely more on top-down or bottom-up signals (Pakan et al., 2018; Han and Helmchen, 2023). Following the same ideas from Bayesian multisensory integration, the brain must hence have mechanisms to track both the reliability of the sensory signals and the predictions thereof (Yon and Frith, 2021).

Here, we show that PE neurons can serve as the backbone for estimating the uncertainty of both the feedforward sensory inputs and the feedback predictions (Figs. 2 & 3). In our model, we assume a hierarchy of PE circuits that are feed-forwardly connected through the lower-order M neuron whose activity encodes the mean of the sensory bottom-up inputs. This local prediction is fed back to the lower-order circuit and at the same time feed-forwarded to the higher-order subnetwork (Fig. 1). With this architecture in place, we show that we rely more strongly on our internal signals when the perceived sensory cues are noisier than the predictions, either because the environment is stable but highly noisy or because the environment changes quickly. This implies that revising our inner model of the world, that is, learning from PEs, should be suppressed if the sensory noise is high or the environment switches rapidly (Herzfeld et al., 2014). Moreover, our work suggests that when experimental trials are short, a neural network can better represent the external world by relying on predictions. Hence, studying neural signatures of predictions in the brain might require experiments that involve sufficiently short trials.

Furthermore, we show that the weighting of sensory inputs and predictions can be biased through neuromodulators, as has been suggested before (see, e.g., Yon and Frith, 2021). In our model, those modulatory signals act through interneurons (Cardin, 2019) whose activities increase in the presence of neuromodulators. When PV neuron activity increases, the network weighs predictions stronger than without modulation. In contrast, when VIP neuron activity increases, the network underestimates the uncertainty of the prediction in a sensory-driven regime, and it underestimates the uncertainty of the sensory input in a prediction-driven regime. Hence, the system leans toward weighting sensory inputs and predictions more equally (Fig. 4A). When SOM and VIP neuron activities are modulated to the same degree, the weighting remains unaffected, suggesting that the individual contributions cancel (Fig. 4B). We show that these findings can be explained by changes in the baseline and gain of PE neurons arising through the modulation of interneuron activity (Fig. 4D). These results can be tested experimentally by optogenetically or pharmacologically stimulating specific interneuron types.

Finally, we illustrate that the weighted integration of feedforward and feedback inputs can be interpreted as a neural manifestation of the contraction bias. We show that the bias is strongly driven by the variances of the sensory cue and the prediction, as well as the trial duration. While the sensory noise increases the contraction bias, the uncertainty in the prediction or long trials decreases the bias (Fig. 5. However, we note that we only consider a neural representation of the stimulus and do not account for other sources of noise, like execution noise, that surely impacts the contraction bias observed in behavioral studies.

Biological evidence for model choices and assumptions

In our model, we assumed that there are dedicated neurons that encode the variance of the feedforward sensory inputs and the prediction. This assumption is consistent with the idea that neurons explicitly encode in their activity the parameters (for instance, mean or variance) of a probability distribution

(O'Neill and Schultz, 2010; O'Reilly et al., 2012). However, how variances of signals are represented in the brain is still not comprehensively understood and alternative ideas have been put forward. For instance, it is conceivable that the variance is encoded in a population of neurons, each differently tuned to a specific parameter (Knill and Pouget, 2004). The neurons' activities represent how close the sensory input is to the preferred (predicted) input of each neuron. Similarly, a neuron's response variability has been suggested to be related to the uncertainty of sensory stimuli (Hoyer and Hyvärinen, 2002; Ma et al., 2006).

There has been evidence that indeed (population of) neurons can encode (un-)certainty (Soltani and Izquierdo, 2019). For instance, neurons in the parietal cortex in monkeys encode the degree of confidence in a perceptual decision (Kiani and Shadlen, 2009). Similarly, the firing rate of neurons in the orbitofrontal cortex have been shown to encode confidence irrespective of sensory modality (Masset et al., 2020). Neural signatures of uncertainty have been found in regions of the prefrontal cortex (Rushworth and Behrens, 2008), the rat insular and orbitofrontal cortex (Jo and Jung, 2016), or the dorsal striatum in monkeys (White and Monosov, 2016). Moreover, the accuracy of memory recalls is encoded in single neurons of the human parietal and temporal lobes Rutishauser et al. (2015, 2018).

In our model, we assume a hierarchy of predictions that are locally computed in memory neurons. These memory neurons are consistent with the idea of internal representation neurons hypothesized in predictive processing theories (Bastos et al., 2012; Keller and Mrsic-Flogel, 2018). While it has been hypothesized that these internal representation neurons might be deeper L5 neurons (Bastos et al., 2012; Heindorf and Keller, 2022), there is also evidence that a group of excitatory L2/3 neurons integrates over negative and positive prediction errors (O'Toole et al., 2022).

The core hypothesis of our model is the presence of sensory PE neurons. Those neurons have been found in different cortical areas in various species (Eliades and Wang, 2008; Keller and Hahnloser, 2009; Ayaz et al., 2019; Audette et al., 2021). Moreover, while first only hypothesized theoretically (Rao and Ballard, 1999), the presence of two types of PE neurons, the negative and positive PE neurons, has been confirmed in several recent studies (Keller et al., 2012; Attinger et al., 2017; Jordan and Keller, 2020; Audette et al., 2021). In our model, the nPE neuron inhibits the memory neuron while the pPE neuron excites it, in line with Keller and Mrsic-Flogel (2018). The weights from those PE neurons onto the M neuron are larger for the lower than the higher PE circuit, so that the lower-order M neuron evolves faster than the higher-order counterpart. This assumption is consistent with the observation that time constants increase along the cortical hierarchy (Murray et al., 2014; Chaudhuri et al., 2015; Runyan et al., 2017).

If the relation of the paces at which the M neurons evolve is strongly modulated, the network either shows a bias towards the sensory cues or the prediction (Figs. S5). It has been hypothesized that symptoms in psychiatric diseases may derive from an erroneous uncertainty estimation (Yon and Frith, 2021). For instance, hallucinations may arise from an underestimation of the expectation uncertainty or an overestimation of the sensory uncertainty. Conversely, a fixation on the environment, even when the sensory cues indicate a switch in the environment, may originate from an overestimation of the expectation uncertainty or an underestimation of the sensory uncertainty (Yon and Frith, 2021).

Neuromodulators and uncertainty

A popular hypothesis is that neuromodulators shape the weighting of sensory inputs and predictions thereof (Yon and Frith, 2021). Theoretical work by (Yu and Dayan, 2005) suggests that acetylcholine (ACh) correlates with *expected uncertainty*, while noradrenaline (NA) correlates with *unexpected uncertainty*. Expected uncertainty is usually interpreted as known cue-outcome unreliabilities. In contrast, unexpected uncertainty relates to the changes in the environment that produce large PEs outside the expected range of uncertainties (Yu and Dayan, 2005). While in our network the stimulus and trail-to-trial variability can only be loosely interpreted as 'expected' and 'unexpected' uncertainty, we want to compare the effects of ACh and NA on the weighting with the effects hypothesized in the literature.

It is assumed that NA increases in more volatile environments and enhances bottom-up processes (Hasselmo et al., 1997; Yon and Frith, 2021). In line with this idea, NA blockade impairs cognitive flexibility (Ridley et al., 1981; Janitzky et al., 2015). In recent work by Lawson et al. (2021), it has been shown that humans receiving propranolol (blocking NA) rely more strongly on their expectations and are slower to update these predictions despite new sensory evidence (Yon and Frith, 2021). A main target for noradrenergic inputs is SOM neurons whose activity increases in the presence of NA (reviewed in, e.g., Urban-Ciecko and Barth, 2016; Hattori et al., 2017; Swanson and Maffei, 2019). In our model, activating SOM neurons does not enhance sensory bottom-up input. In a volatile environment, that is, a sensory-driven regime, the system takes into account predictions slightly more than without SOM

modulation (Figs. 4A and S6).

However, we note that in our simulations, we assumed that neuromodulators act globally, that is, on the interneurons in both the lower and the higher PE circuit. While this agrees with the view that neuromodulators can control network states globally, there is also evidence that they can have a more local, finely adjusted impact on neural circuits (Nadim and Bucher, 2014). In our model, increasing SOM activity only in the lower-order circuit shows a slight enhancement of the sensory weight (Fig. S7), that is, the bottom-up inputs. This suggests that whether a neuromodulator biases the network toward feedforward bottom-up or feedback top-down inputs depends on its spatial and temporal scale of influence.

Similarly to NA, ACh has also been shown to enhance bottom-up, feedforward inputs (reviewed in, e.g., Yu and Dayan, 2005; Marshall et al., 2016). For instance, subjects relied more strongly on prior beliefs when given cholinergic receptor antagonists (Marshall et al., 2016). A major target for cholinergic inputs is VIP neurons whose activity increases in the presence of ACh (reviewed in, e.g., Wester and McBain, 2014; Hattori et al., 2017; Swanson and Maffei, 2019). In our model, activating VIP neurons globally only enhances bottom-up input in stable environments for noisy stimuli. However, increasing VIP activity only in the higher-order PE circuit generally enhances sensory bottom-up inputs (Fig. S7).

Limitations & future steps

As for any computational model, we brush over several biological details to keep the model simple and interpretable. However, those details, while beyond the scope of this study, may be well investigated in future work. For instance, neuromodulatory systems have been suggested to gate plasticity (Pawlak et al., 2010). In a recent study by Jordan and Keller (2023), locus coeruleus (LC) axon activity is shown to correlate with the magnitude of unsigned visuomotor prediction errors. The authors hypothesize that LC output modulates the learning rate at which the internal model evolves (Jordan and Keller, 2023). In our model, we do not consider precision-weighted PEs (but see Wilmes et al., 2023). Hence, a sensible extension to our work would be to adjust the weights from PE neurons onto the M neurons by a function of the stimulus and prediction uncertainty, respectively. This would allow us to compare our results more closely to work showing that ACh and NA can adjust the rate at which new sensory evidence is incorporated when environments change (Marshall et al., 2016; Bruckner et al., 2022).

Our model suggests *one* potential neuronal circuit mechanism for the estimation of sensory inputs and predictions. However, in the light of evidence showing that the integration of feedforward and feedback inputs is species- and modality-dependent, it is conceivable that a plethora of neural mechanisms are used in neural circuits. For multisensory integration, Wong et al. (2023) showed that in Drosophila larva the chosen cue-combination strategy varies depending on the type of sensory information available. Also, humans put typically more weight on visual than auditory cues (Battaglia et al., 2003; Alais and Burr, 2004), but trust vestibular information more than visual information about head direction (Butler et al., 2010), a finding also observed for monkeys (Fetsch et al., 2009). Moreover, Summerfield et al. (2011) showed that humans diverge from an optimal Bayesian strategy in very volatile environments and act according to their experience in the last trial. It has been suggested that the brain may use different strategies to combine signals depending on the task demands (O'Reilly et al., 2012). While these behavioral results cannot speak to the underlying circuit mechanisms, it is conceivable that neural implementations for the integration of feedforward and feedback inputs may also vary.

Furthermore, while we provide a neuronal circuit model for estimating the mean and variance of both sensory signals and predictions, we do not explicitly model the weighting of inputs. A respective neural circuit model would require nested inhibitory interneurons providing divisive inhibition. How this subnetwork interacts with the PE circuits, and how the presence of neuromodulators acting on the interneurons directly involved in the weighting, impacts our findings is subject to future work.

Relation to other work & conclusions

Many normative models have been proposed for state estimation and prediction under uncertainty (Soltani and Izquierdo, 2019), ranging from the classical Kalman filter to more recent models like *Bayes Factor Surprise* (Liakoni et al., 2021). For instance, the Bayes factor surprise formularizes the trade-off between integrating new observations in an existing belief system and resetting this belief system with novel evidence. The surprise factor captures how much an animal's current belief deviates from the new observation.

In recent years, normative models have been squared with biological constraints. For instance, in a

paper by Kutschireiter et al. (2023), uncertainty could be integrated into a ring attractor model encoding head direction. This *Bayesian ring attractor* model encodes uncertainty in the amplitude of the network activity and matches the performance of a circular Kalman filter when the recurrent connections are tuned appropriately. In other seminal work, it has been proposed that Bayesian inference in time can be linked to the dynamics of leaky integrate-and-fire neurons with spike-dependent adaptation (Deneve, 2008).

Here, we proposed an alternative view in which PE neurons serve as the backbone for estimating both the uncertainty of the feedforward sensory stimuli arising from the external world and the feedback signals carrying predictions about the same feedforward inputs our brains are bombarded with. Our work is an important step toward a better understanding of the brain's ability to integrate these often unreliable feedforward and feedback signals that often do not match perfectly.

Models and methods

Network model

The mean-field network model consists of a *lower* and *higher* PE circuit (Fig. 1C-D). Each PE circuit contains an excitatory nPE neuron and pPE neuron ($\mathrm{N_{nPE}} = \mathrm{N_{pPE}} = 1$), as well as inhibitory neurons. The inhibitory neurons comprise PV, SOM and VIP neurons ($\mathrm{N_{SOM}} = \mathrm{N_{VIP}} = 1$, $\mathrm{N_{PV}} = 2$). In addition to the core PE circuit, each subnetwork also includes one memory neuron M and one variance neuron V.

The excitatory neurons in the PE circuit are simulated as two coupled point compartments, representing the soma and the dendrites of elongated pyramidal cells. All other neurons are modeled as point neurons. The activities of all neurons are represented by a set of differential equations describing the network dynamics.

The dynamics of the neurons in the lower and higher PE circuits ($\underline{r}_{\mathrm{PE}}^{\mathrm{low}}$ and $\underline{r}_{\mathrm{PE}}^{\mathrm{high}}$) are given by

$$\begin{aligned} \underline{r}_{\mathrm{PE}}^{\mathrm{low}} &= \left[\underline{h}_{\mathrm{PE}}^{\mathrm{low}}\right]_+ \\ \underline{r}_{\mathrm{PE}}^{\mathrm{high}} &= \left[\underline{h}_{\mathrm{PE}}^{\mathrm{high}}\right]_+ \end{aligned} \tag{1}$$

with

$$\begin{aligned} T \cdot \dot{\underline{h}}_{\mathrm{PE}}^{\mathrm{low}} &= -\underline{h}_{\mathrm{PE}}^{\mathrm{low}} + W_{\mathrm{PE}\leftarrow\mathrm{PE}} \cdot \underline{r}_{\mathrm{PE}}^{\mathrm{low}} + \underline{w}_{\mathrm{PE}\leftarrow\mathrm{M}} \cdot r_{\mathrm{M}}^{\mathrm{low}} + \underline{w}_{\mathrm{PE}\leftarrow\mathrm{FF}} \cdot s + \underline{I}_{\mathrm{PE}} \\ T \cdot \dot{\underline{h}}_{\mathrm{PE}}^{\mathrm{high}} &= -\underline{h}_{\mathrm{PE}}^{\mathrm{high}} + W_{\mathrm{PE}\leftarrow\mathrm{PE}} \cdot \underline{r}_{\mathrm{PE}}^{\mathrm{high}} + \underline{w}_{\mathrm{PE}\leftarrow\mathrm{M}} \cdot r_{\mathrm{M}}^{\mathrm{high}} + \underline{w}_{\mathrm{PE}\leftarrow\mathrm{FF}} \cdot r_{\mathrm{M}}^{\mathrm{low}} + \underline{I}_{\mathrm{PE}}. \end{aligned} \tag{2}$$

We follow the notation that column and row vectors are indicated by letters with an underscore $\underline{\bullet}$, matrices are denoted by capital letters, and scalars are given by small letters without an underscore. Furthermore, a time derivative (e.g., $\frac{dx}{dt}$) is denoted by a dot above the letter (e.g., $\dot{x}$). The rate vector $\underline{r}_{\mathrm{PE}}^{\mathrm{loc}} = \left[r_{\mathrm{nE}}^{\mathrm{loc}},\ r_{\mathrm{pE}}^{\mathrm{loc}},\ r_{\mathrm{nD}}^{\mathrm{loc}},\ r_{\mathrm{pD}}^{\mathrm{loc}},\ r_{\mathrm{PV}_1}^{\mathrm{loc}}, r_{\mathrm{PV}_2}^{\mathrm{loc}},\ r_{\mathrm{SOM}}^{\mathrm{loc}}, r_{\mathrm{VIP}}^{\mathrm{loc}}\right]$ with $\mathrm{loc} \in [\mathrm{low}, \mathrm{high}]$ contains the activities of all neurons or compartments in the PE circuit (soma of nPE/pPE neurons: nE/pE, dendrites of nPE/pPE neurons: nD/pD). The network receives time-dependent stimuli s and neuron/compartment-specific external background input $\underline{I}_{\mathrm{PE}}$. The connection strengths between the *pre*-synaptic neuron (population) and the neurons of the PE circuit are denoted by $W_{\mathrm{PE}\leftarrow\mathrm{pre}}$ and $\underline{w}_{\mathrm{PE}\leftarrow pre}$, respectively. The activities of the neurons evolve with time constants summarized in T.

The activities of the lower and higher M neuron evolve according to a perfect integrator. The memory neurons receive synapses from both nPE and pPE neurons of the same subnetwork,

$$\begin{aligned} \dot{r}_{\mathrm{M}}^{\mathrm{low}} &= \underline{w}_{\mathrm{M}\leftarrow\mathrm{PE}}^{\mathrm{low}} \cdot \underline{r}_{\mathrm{PE}}^{\mathrm{low}} = w_{\mathrm{M}\leftarrow\mathrm{pPE}}^{\mathrm{low}} \cdot r_{\mathrm{pPE}}^{\mathrm{low}} - w_{\mathrm{M}\leftarrow\mathrm{nPE}}^{\mathrm{low}} \cdot r_{\mathrm{nPE}}^{\mathrm{low}} \\ \dot{r}_{\mathrm{M}}^{\mathrm{high}} &= \underline{w}_{\mathrm{M}\leftarrow\mathrm{PE}}^{\mathrm{high}} \cdot \underline{r}_{\mathrm{PE}}^{\mathrm{high}} = w_{\mathrm{M}\leftarrow\mathrm{pPE}}^{\mathrm{high}} \cdot r_{\mathrm{pPE}}^{\mathrm{high}} - w_{\mathrm{M}\leftarrow\mathrm{nPE}}^{\mathrm{high}} \cdot r_{\mathrm{nPE}}^{\mathrm{high}}. \end{aligned} \tag{3}$$

The activities of the lower and higher V neuron evolve according to a leaky integrator with quadratic activation function. The variance neurons receive synapses from both nPE and pPE neurons of the same subnetwork,

$$\begin{aligned} \tau_{\mathrm{V}}^{\mathrm{low}} \cdot \dot{r}_{\mathrm{V}}^{\mathrm{low}} &= -r_{\mathrm{V}}^{\mathrm{low}} + \left(\underline{w}_{\mathrm{V}\leftarrow\mathrm{PE}} \cdot \underline{r}_{\mathrm{PE}}^{\mathrm{low}}\right)^2 = -r_{\mathrm{V}}^{\mathrm{low}} + \left(w_{\mathrm{V}\leftarrow\mathrm{pPE}} \cdot r_{\mathrm{pPE}}^{\mathrm{low}} + w_{\mathrm{V}\leftarrow\mathrm{nPE}} \cdot r_{\mathrm{nPE}}^{\mathrm{low}}\right)^2 \\ \tau_{\mathrm{V}}^{\mathrm{high}} \cdot \dot{r}_{\mathrm{V}}^{\mathrm{high}} &= -r_{\mathrm{V}}^{\mathrm{high}} + \left(\underline{w}_{\mathrm{V}\leftarrow\mathrm{PE}} \cdot \underline{r}_{\mathrm{PE}}^{\mathrm{high}}\right)^2 = -r_{\mathrm{V}}^{\mathrm{high}} + \left(w_{\mathrm{V}\leftarrow\mathrm{pPE}} \cdot r_{\mathrm{pPE}}^{\mathrm{high}} + w_{\mathrm{V}\leftarrow\mathrm{nPE}} \cdot r_{\mathrm{nPE}}^{\mathrm{high}}\right)^2. \end{aligned} \tag{4}$$

All values for neuron and network parameters, details on the model equations for the mean-field and the population network, as well as supporting analyses can be found in the supplementary material.

Weighting of sensory inputs and predictions

We arithmetically calculated the weighted output of sensory inputs and predictions, r_{out}, based on ideas of Bayesian multisensory integration (see, e.g. Pouget et al., 2013),

$$r_{\text{out}} = \alpha \cdot s + (1 - \alpha) \cdot r_{\text{M}}^{\text{low}}, \tag{5}$$

where α denotes the sensory weight (that is, the reliability of the sensory input) and is given by

$$\alpha = \left(1 + \frac{r_{\text{V}}^{\text{low}}}{r_{\text{V}}^{\text{high}}}\right)^{-1}. \tag{6}$$

Inputs

The network receives feedforward stimuli s that may vary between trials. To account for noise, each stimulus is composed of N_{in} consecutive, piece-wise constant values drawn from a normal distribution with mean μ_{in} and standard deviation σ_{in}. To account for changes in the environment, μ_{in} is drawn from a uniform distribution $U(a, b)$ with mean $\mu_{\text{trial}} = \frac{a+b}{2}$ and standard deviation $\sigma_{\text{trial}} = \frac{b-a}{\sqrt{12}}$. The parameterization of both distributions varies across the experiments. All stimulus/input parameters can be found in the supplementary material.

Simulations

All simulations were performed in customized Python code written by LH. Source code to reproduce the simulations, analyses, and figures will be available after publication at `https://github.com/lhertaeg/weighted_sensory_prediction`. Differential equations were numerically integrated using a 2^{nd}-order Runge-Kutta method. Neurons were initialized with $r = 0/s$. Further details and values for simulation parameters can be found in the supplementary material.

Acknowledgments

References

Athena Akrami, Charles D Kopec, Mathew E Diamond, and Carlos D Brody. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature*, 554(7692):368–372, 2018.

David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004.

Paymon Ashourian and Yonatan Loewenstein. Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS one*, 6(5):e19551, 2011.

Alexander Attinger, Bo Wang, and Georg B Keller. Visuomotor coupling shapes the functional development of mouse visual cortex. *Cell*, 169(7):1291–1302, 2017.

Nicholas J Audette, WenXi Zhou, and David M Schneider. Temporally precise movement-based predictions in the mouse auditory cortex. *bioRxiv*, pages 2021–12, 2021.

Michael C Avery and Jeffrey L Krichmar. Neuromodulatory systems and their interactions: a review of models, theories, and experiments. *Frontiers in neural circuits*, 11:108, 2017.

Aslı Ayaz, Andreas Stäuble, Morio Hamada, Marie-Angela Wulf, Aman B Saleem, and Fritjof Helmchen. Layer-specific integration of locomotion and sensory information in mouse barrel cortex. *Nature communications*, 10(1):2585, 2019.

Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

Peter W Battaglia, Robert A Jacobs, and Richard N Aslin. Bayesian integration of visual and auditory signals for spatial localization. *Josa a*, 20(7):1391–1397, 2003.

Rasmus Bruckner, Hauke R Heekeren, and Matt Nassar. Understanding learning through uncertainty and bias. 2022.

John S Butler, Stuart T Smith, Jennifer L Campos, and Heinrich H Bülthoff. Bayesian integration of visual and vestibular signals for heading. *Journal of vision*, 10(11):23–23, 2010.

Luke Campagnola, Stephanie C Seeman, Thomas Chartrand, Lisa Kim, Alex Hoggarth, Clare Gamlin, Shinya Ito, Jessica Trinh, Pasha Davoudian, Cristina Radaelli, et al. Local connectivity and synaptic dynamics in mouse and human neocortex. *Science*, 375(6585):eabj5861, 2022.

Jessica A Cardin. Functional flexibility in cortical circuits. *Current opinion in neurobiology*, 58:175–180, 2019.

Rishidev Chaudhuri, Kenneth Knoblauch, Marie-Alice Gariel, Henry Kennedy, and Xiao-Jing Wang. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron*, 88 (2):419–431, 2015.

Sophie Deneve. Bayesian spiking neurons i: inference. *Neural computation*, 20(1):91–117, 2008.

Sophie Deneve and Alexandre Pouget. Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology-Paris*, 98(1-3):249–258, 2004.

Steven J Eliades and Xiaoqin Wang. Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature*, 453(7198):1102, 2008.

Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Christopher R Fetsch, Amanda H Turner, Gregory C DeAngelis, and Dora E Angelaki. Dynamic reweighting of visual and vestibular cues during self-motion perception. *Journal of Neuroscience*, 29 (49):15601–15612, 2009.

Christopher R Fetsch, Alexandre Pouget, Gregory C DeAngelis, and Dora E Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature neuroscience*, 15(1):146–154, 2012.

Karl Friston. Hierarchical models in the brain. *PLoS computational biology*, 4(11):e1000211, 2008.

Yong Gu, Dora E Angelaki, and Gregory C DeAngelis. Neural correlates of multisensory cue integration in macaque mstd. *Nature neuroscience*, 11(10):1201–1210, 2008.

Shuting Han and Fritjof Helmchen. Behavior-relevant top-down cross-modal predictions in mouse neocortex. *bioRxiv*, pages 2023–04, 2023.

Kenneth D Harris and Gordon MG Shepherd. The neocortical circuit: themes and variations. *Nature neuroscience*, 18(2):170, 2015.

Michael E Hasselmo, Christiane Linster, Madhvi Patil, Daveena Ma, and Milos Cekic. Noradrenergic suppression of synaptic transmission may influence cortical signal-to-noise ratio. *Journal of neurophysiology*, 77(6):3326–3339, 1997.

Ryoma Hattori, Kishore V Kuchibhotla, Robert C Froemke, and Takaki Komiyama. Functions and dysfunctions of neocortical inhibitory neuron subtypes. *Nature neuroscience*, 20(9):1199–1208, 2017.

Matthias Heindorf and Georg B Keller. Reduction of layer 5 mediated long-range cortical communication by antipsychotic drugs. *bioRxiv*, 2022.

Loreen Hertäg and Claudia Clopath. Prediction-error neurons in circuits with multiple neuron types: Formation, refinement, and functional implications. *Proceedings of the National Academy of Sciences*, 119(13):e2115699119, 2022.

Loreen Hertäg and Henning Sprekeler. Learning prediction error neurons in a canonical interneuron circuit. *Elife*, 9:e57541, 2020.

David J Herzfeld, Pavan A Vaswani, Mollie K Marko, and Reza Shadmehr. A memory of errors in sensorimotor learning. *Science*, 345(6202):1349–1353, 2014.

Harry Levi Hollingworth. The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17):461–469, 1910.

Patrik Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. *Advances in neural information processing systems*, 15, 2002.

Kathrin Janitzky, Michael T Lippert, Achim Engelhorn, Jennifer Tegtmeier, Jürgen Goldschmidt, Hans-Jochen Heinze, and Frank W Ohl. Optogenetic silencing of locus coeruleus activity in mice impairs cognitive flexibility in an attentional set-shifting task. *Frontiers in behavioral neuroscience*, 9:286, 2015.

Mehrdad Jazayeri and Michael N Shadlen. Temporal context calibrates interval timing. *Nature neuroscience*, 13(8):1020–1026, 2010.

Xiaolong Jiang, Shan Shen, Cathryn R Cadwell, Philipp Berens, Fabian Sinz, Alexander S Ecker, Saumil Patel, and Andreas S Tolias. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, 350(6264):aac9462, 2015.

Suhyun Jo and Min Whan Jung. Differential coding of uncertain reward in rat insular and orbitofrontal cortex. *Scientific reports*, 6(1):24085, 2016.

Rebecca Jordan and Georg B Keller. Opposing influence of top-down and bottom-up input on excitatory layer 2/3 neurons in mouse primary visual cortex. *Neuron*, 108(6):1194–1206, 2020.

Rebecca Jordan and Georg B Keller. The locus coeruleus broadcasts prediction errors across the cortex to promote sensorimotor plasticity. *Elife*, 12:RP85111, 2023.

Georg B Keller and Richard HR Hahnloser. Neural processing of auditory feedback during vocal practice in a songbird. *Nature*, 457(7226):187, 2009.

Georg B Keller and Thomas D Mrsic-Flogel. Predictive processing: A canonical cortical computation. *Neuron*, 100(2):424–435, 2018.

Georg B Keller, Tobias Bonhoeffer, and Mark Hübener. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, 74(5):809–815, 2012.

Roozbeh Kiani and Michael N Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, 324(5928):759–764, 2009.

David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427 (6971):244–247, 2004.

Anna Kutschireiter, Melanie A Basnak, Rachel I Wilson, and Jan Drugowitsch. Bayesian inference in ring attractor networks. *Proceedings of the National Academy of Sciences*, 120(9):e2210622120, 2023.

Matthew Larkum. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in neurosciences*, 36(3):141–151, 2013.

Rebecca P Lawson, James Bisby, Camilla L Nord, Neil Burgess, and Geraint Rees. The computational, pharmacological, and physiological determinants of sensory learning under uncertainty. *Current Biology*, 31(1):163–172, 2021.

Vasiliki Liakoni, Alireza Modirshanechi, Wulfram Gerstner, and Johanni Brea. Learning in volatile environments with the bayes factor surprise. *Neural Computation*, 33(2):269–340, 2021.

Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*, 5(10):793, 2004.

Louise Marshall, Christoph Mathys, Diane Ruge, Archy O De Berker, Peter Dayan, Klaas E Stephan, and Sven Bestmann. Pharmacological fingerprints of contextual uncertainty. *PLoS Biology*, 14(11): e1002575, 2016.

Paul Masset, Torben Ott, Armin Lak, Junya Hirokawa, and Adam Kepecs. Behavior-and modality-general representation of confidence in orbitofrontal cortex. *Cell*, 182(1):112–126, 2020.

Nicolas Meirhaeghe, Hansem Sohn, and Mehrdad Jazayeri. A precise and adaptive neural mechanism for predictive temporal processing in the frontal cortex. 109(18):2995–3011.e5. ISSN 0896-6273. doi: 10.1016/j.neuron.2021.08.025. URL `https://doi.org/10.1016/j.neuron.2021.08.025`. Publisher: Elsevier.

David Mumford. On the computational architecture of the neocortex. *Biological cybernetics*, 66(3): 241–251, 1992.

John D Murray, Alberto Bernacchia, David J Freedman, Ranulfo Romo, Jonathan D Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, et al. A hierarchy of intrinsic timescales across primate cortex. *Nature neuroscience*, 17(12):1661–1663, 2014.

Farzan Nadim and Dirk Bucher. Neuromodulation of neurons and synapses. *Current opinion in neurobiology*, 29:48–56, 2014.

Martin O'Neill and Wolfram Schultz. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron*, 68(4):789–800, 2010.

Jill X O'Reilly, Saad Jbabdi, and Timothy EJ Behrens. How can a bayesian approach inform neuroscience? *European Journal of Neuroscience*, 35(7):1169–1179, 2012.

Sean M O'Toole, Hassana K Oyibo, and Georg B Keller. Prediction error neurons in mouse cortex are molecularly targetable cell types. *BioRxiv*, pages 2022–07, 2022.

Janelle MP Pakan, Stephen P Currie, Lukas Fischer, and Nathalie L Rochefort. The impact of visual cues, reward, and motor feedback on the representation of behaviorally relevant spatial locations in primary visual cortex. *Cell reports*, 24(10):2521–2528, 2018.

Verena Pawlak, Jeffery R Wickens, Alfredo Kirkwood, and Jason ND Kerr. Timing is not everything: neuromodulation opens the stdp gate. *Frontiers in synaptic neuroscience*, 2:146, 2010.

Frederike H Petzschner and Stefan Glasauer. Iterative bayesian estimation as an explanation for range and regression effects: a study on human path integration. *Journal of Neuroscience*, 31(47):17220–17229, 2011.

Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature neuroscience*, 16(8):1068–1076, 2013.

Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–1178, 2013.

Brian C Rakitin, John Gibbon, Trevor B Penney, Chara Malapani, Sean C Hinton, and Warren H Meck. Scalar expectancy theory and peak-interval timing in humans. *Journal of Experimental Psychology: Animal Behavior Processes*, 24(1):15, 1998.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79, 1999.

RM Ridley, TAJ Haystead, HF Baker, and TJ Crow. A new approach to the role of noradrenaline in learning: problem-solving in the marmoset after α-noradrenergic receptor blockade. *Pharmacology Biochemistry and Behavior*, 14(6):849–855, 1981.

Benjamin Rowland, Terrence Stanford, and Barry Stein. A bayesian model unifies multisensory spatial localization with the physiological properties of the superior colliculus. *Experimental Brain Research*, 180:153–161, 2007.

Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neurobiology*, 71(1):45–61, 2011.

Caroline A Runyan, Eugenio Piasini, Stefano Panzeri, and Christopher D Harvey. Distinct timescales of population coding across cortex. *Nature*, 548(7665):92–96, 2017.

Matthew FS Rushworth and Timothy EJ Behrens. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature neuroscience*, 11(4):389–397, 2008.

Ueli Rutishauser, Shengxuan Ye, Matthieu Koroma, Oana Tudusciuc, Ian B Ross, Jeffrey M Chung, and Adam N Mamelak. Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nature neuroscience*, 18(7):1041–1050, 2015.

Ueli Rutishauser, Tyson Aflalo, Emily R Rosario, Nader Pouratian, and Richard A Andersen. Single-neuron representation of memory strength and recognition confidence in left human posterior parietal cortex. *Neuron*, 97(1):209–220, 2018.

Alireza Soltani and Alicia Izquierdo. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, 20(10):635–644, 2019.

Christopher Summerfield, Timothy E Behrens, and Etienne Koechlin. Perceptual classification in a rapidly changing environment. *Neuron*, 71(4):725–736, 2011.

Olivia K Swanson and Arianna Maffei. From hiring to firing: activation of inhibitory neurons and their recruitment in behavior. *Frontiers in molecular neuroscience*, 12:168, 2019.

Robin Tremblay, Soohyun Lee, and Bernardo Rudy. Gabaergic interneurons in the neocortex: from cellular properties to circuits. *Neuron*, 91(2):260–292, 2016.

Joanna Urban-Ciecko and Alison L Barth. Somatostatin-expressing neurons in cortical networks. *Nature Reviews Neuroscience*, 17(7):401–409, 2016.

Mark T Wallace, M Alex Meredith, and Barry E Stein. Multisensory integration in the superior colliculus of the alert cat. *Journal of neurophysiology*, 80(2):1006–1010, 1998.

Jason C Wester and Chris J McBain. Behavioral state-dependent modulation of distinct interneuron subtypes and consequences for circuit function. *Current opinion in neurobiology*, 29:118–125, 2014.

J Kael White and Ilya E Monosov. Neurons in the primate dorsal striatum signal the uncertainty of object–reward associations. *Nature communications*, 7(1):12735, 2016.

Katharina A Wilmes, Mihai A Petrovici, Shankar Sachidhanandam, and Walter Senn. Uncertainty-modulated prediction errors in cortical microcircuits. *bioRxiv*, pages 2023–05, 2023.

Hugh R Wilson and Jack D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.

Philip Wong, Andreas Braun, Daniel Malagarriga, Jeff Moehlis, Ruben Moreno Bote, Alexandre Pouget, and Matthieu Louis. Computational principles of adaptive multisensory combination in the drosophila larva. *bioRxiv*, pages 2023–05, 2023.

Daniel Yon and Chris D Frith. Precision and the bayesian brain. *Current Biology*, 31(17):R1026–R1032, 2021.

Angela J Yu and Peter Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692, 2005.

Supporting Information

A Detailed Methods

In the following, we describe in more detail the equations for the dynamics of the neurons in the prediction-error circuit, as well as the memory and variance neurons. We then provide the connectivity of the network and the inputs to the neurons for both the mean-field and population model. Finally, to ensure reproducibility, we summarize all simulation parameters used for the results shown in the figures.

A.1 Network model

The network model consists of a *lower* and *higher* mean-field PE circuit (Fig. 1). Each PE circuit contains an excitatory nPE neuron and pPE neuron ($N_{nPE} = N_{pPE} = 1$), as well as inhibitory neurons. The inhibitory neurons comprise PV, SOM and VIP neurons ($N_{SOM} = N_{VIP} = 1$, $N_{PV} = 2$). In addition to the core PE circuit, each subnetwork also includes one memory neuron M and one variance neuron V.

In Figure 2 and the corresponding supporting figures, only the lower subnetwork is simulated. In Fig. S2, we replaced this lower mean-field PE circuit with a heterogeneous population model containing 200 neurons ($N_{SOM} = N_{VIP} = N_{PV} = 20$, 140 excitatory neurons). In Fig. S3, the lower PE circuit comprises 1000 copies of the mean-field network to account for selectivity.

In the following, we describe the dynamics of the neurons/compartments in the mean-field network. The equations for the population PE circuit (Fig. S2) are directly deduced from the mean-field equations and can also be found in (Hertäg and Clopath, 2022).

A.1.1 Prediction-error network model

Each excitatory pyramidal cell (that is, nPE or pPE neuron) is divided into two coupled compartments, representing the soma and the dendrites, respectively. The dynamics of the firing rates of the somatic compartments r_{nE} (nPE neuron) and r_{pE} (pPE neuron) obey (Wilson and Cowan, 1972)

$$\begin{aligned} r_{nE} &= [h_{nE}]_+ \quad \text{with} \quad \tau_E \frac{dh_{nE}}{dt} = -h_{nE} + w_{nE\leftarrow nD} \cdot r_{nD} - w_{nE\leftarrow PV_1} \cdot r_{PV_1} - w_{nE\leftarrow PV_2} \cdot r_{PV_2} + I_{nE}, \\ r_{pE} &= [h_{pE}]_+ \quad \text{with} \quad \tau_E \frac{dh_{pE}}{dt} = -h_{pE} + w_{pE\leftarrow pD} \cdot r_{pD} - w_{pE\leftarrow PV_1} \cdot r_{PV_1} - w_{pE\leftarrow PV_2} \cdot r_{PV_2} + I_{pE} \end{aligned} \quad (7)$$

where τ_E denotes the excitatory rate time constant (τ_E=60 ms), the weights $w_{nE\leftarrow nD}$ and $w_{pE\leftarrow pD}$ describe the connection strength between the dendritic compartment and the soma of the same neuron, and $w_{nE\leftarrow PV_1}$, $w_{nE\leftarrow PV_2}$, $w_{pE\leftarrow PV_1}$ and $w_{pE\leftarrow PV_2}$ denote the strength of somatic inhibition from PV neurons.

The overall input I_{nE} and I_{pE} comprise the external background and feedforward inputs (see "Inputs" below). Firing rates are rectified to ensure positivity ($[\bullet]_+$).

The dynamics of the activity of the dendritic compartments r_{nD} (nPE neuron) and r_{pD} (pPE neuron) obey (Wilson and Cowan, 1972)

$$\begin{aligned} r_{\mathrm{nD}} = [h_{\mathrm{nD}}]_+ \quad \text{with} \quad \tau_E \, \frac{dh_{\mathrm{nD}}}{dt} &= -h_{\mathrm{nD}} + w_{\mathrm{nD}\leftarrow\mathrm{nE}} \cdot r_{\mathrm{nE}} + w_{\mathrm{nD}\leftarrow\mathrm{pE}} \cdot r_{\mathrm{pE}} + w_{\mathrm{nD}\leftarrow\mathrm{M}} \cdot r_{\mathrm{M}} \\ &\quad - w_{\mathrm{nD}\leftarrow\mathrm{SOM}} \cdot r_{\mathrm{SOM}} + I_{\mathrm{nD}}, \\ r_{\mathrm{pD}} = [h_{\mathrm{pD}}]_+ \quad \text{with} \quad \tau_E \, \frac{dh_{\mathrm{pD}}}{dt} &= -h_{\mathrm{pD}} + w_{\mathrm{pD}\leftarrow\mathrm{nE}} \cdot r_{\mathrm{pE}} + w_{\mathrm{pD}\leftarrow\mathrm{pE}} \cdot r_{\mathrm{pE}} + w_{\mathrm{pD}\leftarrow\mathrm{M}} \cdot r_{\mathrm{M}} \\ &\quad - w_{\mathrm{pD}\leftarrow\mathrm{SOM}} \cdot r_{\mathrm{SOM}} + I_{\mathrm{pD}}, \end{aligned} \tag{8}$$

where the weights $w_{\mathrm{nD}\leftarrow\mathrm{nE}}$, $w_{\mathrm{nD}\leftarrow\mathrm{pE}}$, $w_{\mathrm{pD}\leftarrow\mathrm{nE}}$ and $w_{\mathrm{pD}\leftarrow\mathrm{pE}}$ denote the recurrent excitatory connections between PCs, including backpropagating activity from the soma to the dendrites. $w_{\mathrm{nD}\leftarrow\mathrm{SOM}}$ and $w_{\mathrm{pD}\leftarrow\mathrm{SOM}}$ represent the strength of dendritic inhibition from the SOM neuron. $w_{\mathrm{nD}\leftarrow\mathrm{M}}$ and $w_{\mathrm{pD}\leftarrow\mathrm{M}}$ denote the strength of connection between the memory neuron and the dendrites. The overall inputs I_{nD} and I_{pD} comprise fixed, external background inputs (see "Inputs" below). We assume that any excess of inhibition in a dendrite does not affect the soma, that is, the dendritic compartment is rectified at zero.

Similarly, the firing rate dynamics of each interneuron is modeled by a rectified, linear differential equation,

$$\begin{aligned} r_{\mathrm{X}} = [h_{\mathrm{X}}]_+ \quad \text{with} \quad \tau_I \, \frac{dh_{\mathrm{X}}}{dt} &= -h_{\mathrm{X}} + I_{\mathrm{X}} + w_{\mathrm{X}\leftarrow\mathrm{nE}} \cdot r_{\mathrm{nE}} + w_{\mathrm{X}\leftarrow\mathrm{pE}} \cdot r_{\mathrm{pE}} + w_{\mathrm{X}\leftarrow\mathrm{M}} \cdot r_{\mathrm{M}} - w_{\mathrm{X}\leftarrow\mathrm{PV}_1} \cdot r_{\mathrm{PV}_1} \\ &\quad - w_{\mathrm{X}\leftarrow\mathrm{PV}_2} \cdot r_{\mathrm{PV}_2} - w_{\mathrm{X}\leftarrow\mathrm{SOM}} \cdot r_{\mathrm{SOM}} - w_{\mathrm{X}\leftarrow\mathrm{VIP}} \cdot r_{\mathrm{VIP}}, \end{aligned} \tag{9}$$

where r_{X} denotes the firing rate of interneuron type X, and the weight $w_{\mathrm{X}\leftarrow\mathrm{Y}}$ denotes the strength of connection between the presynaptic neuron Y and the postsynaptic neuron X ($X \in \{\mathrm{PV}_1, \mathrm{PV}_2, \mathrm{SOM}, \mathrm{VIP}\}$, $Y \in \{\mathrm{nPE}, \mathrm{pPE}, \mathrm{PV}_1, \mathrm{PV}_2, \mathrm{SOM}, \mathrm{VIP}, \mathrm{M}\}$). The rate time constant τ_I was chosen to resemble a fast $\mathrm{GABA_A}$ time constant, and set to 2 ms for all interneuron types included. The overall input I_{X} comprises fixed, external background inputs and feedforward sensory inputs (see "Inputs" below).

A.1.2 Memory and variance neuron

In addition to the core PE circuit, we simulate a memory neuron M and a variance neuron V. The memory neuron is modeled as a perfect integrator, receiving synapses from both the nPE and pPE neuron,

$$\tau_E \cdot \frac{dr_{\mathrm{M}}}{dt} = w_{\mathrm{M}\leftarrow\mathrm{pE}} \cdot r_{\mathrm{pE}} - w_{\mathrm{M}\leftarrow\mathrm{nE}} \cdot r_{\mathrm{nE}}. \tag{10}$$

$w_{\mathrm{M}\leftarrow\mathrm{pE}}$ denotes the connection strength between the pPE neuron and the memory neuron, and $w_{\mathrm{M}\leftarrow\mathrm{nE}}$ denotes the connection strength between the nPE neuron and the memory neuron. The time constant $\tau_E = 60$ ms.

The dynamics of the variance neuron obeys a non-linear differential equation with leak term,

$$\tau_V \cdot \frac{dr_{\mathrm{V}}}{dt} = -r_{\mathrm{V}} + (w_{\mathrm{V}\leftarrow\mathrm{pE}} \cdot r_{\mathrm{pE}} + w_{\mathrm{V}\leftarrow\mathrm{nE}} \cdot r_{\mathrm{nE}})^2. \tag{11}$$

The weight $w_{\mathrm{V}\leftarrow\mathrm{pE}}$ represents the connection strength between the pPE neuron and the variance neuron, while $w_{\mathrm{V}\leftarrow\mathrm{nE}}$ denotes the connection strength between the nPE neuron and the variance neuron. To ensure that the V neuron encodes the variance, we chose a quadratic activation function. In Fig. S5, we used a linear activation function to investigate the impact of the input-output transfer function on the weighting of sensory inputs and predictions. The time constant τ_V was 5 s in the mean-field model, 2 s in the heterogeneous population model (Fig. S2), and 0.5 s in the population model with selectivity (Fig. S3).

A.1.3 Weighted output

The weighted output $r_{\mathrm{out}}(t)$ is a linear combination of the current sensory input $s(t)$ and the activity of the memory neuron, $r_{\mathrm{M}}(t)$, inspired by Bayesian multisensory integration (see, e.g. Pouget et al., 2013),

$$r_{\mathrm{out}}(t) = \alpha \cdot s(t) + (1 - \alpha) \cdot r_{\mathrm{M}}(t). \tag{12}$$

How strongly either the sensory input or the prediction thereof contributes to the output is denoted by the sensory weight α,

$$\begin{aligned}\alpha &= \frac{r_{\mathrm{V_{lower}}}^{-1}}{r_{\mathrm{V_{lower}}}^{-1} + r_{\mathrm{V_{higher}}}^{-1}} \\ &= \left(1 + \frac{r_{\mathrm{V_{lower}}}}{r_{\mathrm{V_{higher}}}}\right)^{-1}. \end{aligned} \tag{13}$$

A.2 Connectivity

A.2.1 Connections between neurons of the PE circuit

The connectivity between neurons of the PE circuit, both for the mean-field and the population network, were taken from (Hertäg and Clopath, 2022). We considered three mean-field networks (see Table 2) that differed in terms of the inputs (feedforward vs. feedback) onto the SOM and VIP neurons, and, hence, in their connectivity that established an E/I balance in the excitatory neurons.

A.2.2 Connections between the PE circuit and the M neuron

While the nPE neurons inhibit the M neuron, the pPE neurons excite it. To ensure that the activities of the memory neurons represent the mean of the sensory stimuli in the lower PE circuit and the mean of the prediction in the higher subnetwork, respectively, the net effect of nPE and pPE neurons must cancel in the steady state (see Analysis in B.2). Hence, the weights need to account for the neurons' potentially different gain factors (g_{nPE} and g_{pPE}) and the neuron numbers (N_{nPE} and N_{pPE}):

$$\begin{aligned} w_{\mathrm{M}\leftarrow\mathrm{nE}} &= \frac{-\lambda}{g_{\mathrm{nPE}} \cdot N_{\mathrm{nPE}}} \\ w_{\mathrm{M}\leftarrow\mathrm{pE}} &= \frac{\lambda}{g_{\mathrm{pPE}} \cdot N_{\mathrm{pPE}}} \end{aligned}$$

where λ denotes the speed at which the perfect integrator evolves. In the lower PE circuit, $\lambda = 3 \cdot 10^{-3}$ for the mean-field model in Fig. 2 and $\lambda = 4.5 \cdot 10^{-2}$ otherwise. In the higher PE circuit, $\lambda = 7 \cdot 10^{-4}$. For the heterogeneous population model, $\lambda = 5 \cdot 10^{-2}$ in Fig. S2, and for the population model with selectivity (Fig. S3), $\lambda = 3 \cdot 10^{-1}$.

For the mean-field networks ($N_{\mathrm{nPE}} = N_{\mathrm{pPE}} = 1$), the gain factors are given in Table 1. For the population network, the gain factors for all PE neurons are shown in Fig. 6.

Network	**MFN 1** FF → SOM FB → VIP	**MFN 2** FB → SOM FF → VIP	**MFN 3** FF → SOM FF → VIP
nPE	1	1.7	2.5
pPE	1	1.7	2.5

Table 1. Gain factors for nPE and pPE neurons in three different mean-field networks (MFN). Each MFN differs with respect to the inputs onto SOM and VIP neurons. The interneurons either receive the feedforward (FF) or feedback (FB) input. All numbers are rounded to the first digit.

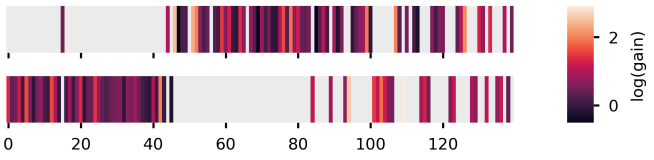

Figure 6. Gain factors of nPE and pPE neurons in the population model.
The logarithm of the gain factors of nPE (top) and pPE (bottom) neurons in the population model from (Hertäg and Clopath, 2022). The network contains 67 nPE neurons and 66 pPE neurons. The remaining excitatory neurons were not classified as PE neurons and were not connected to the M neuron.

The memory neuron M connects to the post-synaptic neurons X in the PE circuit with the connection strength $w_{\mathrm{X}\leftarrow\mathrm{M}}$. If a connection exists, we assume $w_{\mathrm{X}\leftarrow\mathrm{M}} = 1$. In all mean-field networks and the

population network, the dendrites of nPE and pPE neurons and one of the two (populations of) PV neurons receive connections from the memory neuron. Furthermore, we assume that the M neuron does not excite the soma of PCs. Whether the SOM or VIP neurons are the target of the feedback projections depend on the specific mean-field network (see Table 2). In the population model, 30% of the SOM neurons and 70% of the VIP neurons receive input from the memory neuron.

Network	**MFN 1** FF → SOM FB → VIP	**MFN 2** FB → SOM FF → VIP	**MFN 3** FF → SOM FF → VIP
SOM	0	1	0
VIP	1	0	0

Table 2. $w_{\mathrm{X}\leftarrow\mathrm{M}}$ for the post-synaptic SOM and VIP neurons in all three mean-field networks considered.

A.2.3 Connections between the PE circuit and the V neuron

Both nPE and pPE neurons excite the V neuron. To ensure that the activity of the V neuron represents the variance of the input (see Analysis in B.2), the weights must account for differences in the gains (g_{nPE} and g_{pPE}, see Table 1 and Fig. 6) and numbers (N_{nPE} and N_{pPE}) of the PE neurons,

$$\begin{aligned} w_{\mathrm{V}\leftarrow\mathrm{nE}} &= \frac{\theta}{g_{\mathrm{nPE}} \cdot N_{\mathrm{nPE}}} \\ w_{\mathrm{V}\leftarrow\mathrm{pE}} &= \frac{\theta}{g_{\mathrm{pPE}} \cdot N_{\mathrm{pPE}}}. \end{aligned}$$

We introduce the factor θ to compensate for cross-terms that are the result of a quadratic activation function but are not in line with the definition of the variance.

In the mean-field network, $\theta = 1$ because nPE and pPE neuron activity is mutually exclusive, and, hence, the cross-term $\mathrm{nPE} \cdot \mathrm{pPE}$ would be zero (under the assumption that they have a negligible baseline activity). This is also true for the population model, in which each nPE neuron only contributes a small fraction to the overall, mean nPE, and each pPE neuron only contributes a small fraction to the overall, mean pPE.

However, in Supp Fig. S3, each mean-field network receives a stimulus s_i drawn from a distribution. In this case, θ must be chosen such that deviations from the true variance can be mitigated or fully corrected. The true θ depends on the distribution at hand. In our simulations, we used a uniform distribution $U(a, b)$, in which case θ can be derived from

$$\begin{aligned} \left(\sum_i r_{\mathrm{nE,i}} + \sum_i r_{\mathrm{pE,i}}\right)^2 &\overset{(i)}{=} \left(\frac{\theta}{N}\sum_{s_i\geq\mu}^{N/2}(s_i-\mu) + \frac{\theta}{N}\sum_{s_i\leq\mu}^{N/2}(\mu-s_i)\right)^2 \\ &= \frac{\theta^2}{4}\left(\frac{2}{N}\sum_{s_i\geq\mu}^{N/2} s_i - \frac{2}{N}\sum_{s_i\leq\mu}^{N/2} s_i\right)^2 \\ &\overset{(ii)}{=} \frac{\theta^2}{4}\left(\frac{b+\mu}{2} - \frac{\mu+a}{2}\right)^2 = \frac{(b-a)^2}{16}\cdot\theta^2, \end{aligned} \tag{14}$$

where we assumed that (i) the number of nPE and pPE neurons is equal, and (ii) this number goes to infinity. Comparing eq. (14) with the equation for the variance of a uniform distribution, $\frac{(b-a)^2}{12}$, we get $\theta = \frac{2}{\sqrt{3}}$.

A.3 Inputs

Each neuron (type) receives an overall input I_i,

$$I_i = I_i^{BL} + w_i \cdot I_i^{FF}$$

where I_i^{FF} denotes a feedforward input and I_i^{BL} represents an external background input that ensures reasonable baseline firing rates in the absence of sensory inputs and predictions thereof. In the case of

the mean-field network, these inputs were set such that the baseline firing rates are $r_{\mathrm{pE}} = r_{\mathrm{pD}} = r_{\mathrm{nE}} = r_{\mathrm{nD}} = 0\,s^{-1}$ and $r_{\mathrm{P}} = r_{\mathrm{S}} = r_{\mathrm{V}} = 4\,s^{-1}$. In the case of the population network, we set the external inputs of all neuron types to $5\,s^{-1}$, while the background inputs to the dendrites were computed such that the dendrites are inactive during baseline.

The feedforward input is either the direct sensory input s for the lower PE circuit, or the activity of the M neuron, r_{M}, for the higher PE circuit. In general, for the three mean-field networks tested, we chose $w_i = 1 - w_{\mathrm{X}\leftarrow\mathrm{M}}$ (see Table 2). In the population network, 70% of the SOM neurons and 30% of the VIP neurons receive the feedforward input.

A.4 Simulations

All simulations were performed in customized Python code written by LH. Source code to reproduce the simulations, analyses and figures will be available after publication at `https://github.com/lhertaeg/weighted_sensory_prediction`. Differential equations were numerically integrated using a 2^{nd}-order Runge-Kutta method. Neurons were initialized with $r = 0/s$.

The qualitative results were fairly robust to the choice of the simulation parameters and are here stated merely to ensure the reproducibility of all figures. However, we note that we made use of PE circuits that had been trained on steady state inputs (Hertäg and Clopath, 2022). Hence, we must simulate the network long enough to ensure that the PE neurons reach their steady state. Moreover, the lower-level M neuron must evolve faster than the higher-level M neuron as indicated in Fig. S5. Finally, the time constant of the V neurons must be of the same magnitude as the trial duration.

In the following, we give all figure-specific parameters not directly visible or mentioned in the figures and captions. Furthermore, to increase readability, we do not include units for the parameters. All units can be deduced from the equations above. We simulated the network in Figures 1 and 2 for 10^5 simulation time steps. In Figure 2, we presented 200 constant values, each 500 time steps long. In Figure 3 (including supporting figures) we simulated 100 trials, while in Figures 4 to 5 (including supporting figures), we simulated 200 trials, each 5000 time steps long. In a trail, 10 constant values were drawn from a normal distribution $N(\mu_{\mathrm{in}}, \sigma^2_{\mathrm{in}})$, each 500 time steps long. The stimulus mean was drawn from an uniform distribution, $U(a, b)$, with mean μ_{in} and variance $\sigma^2_{\mathrm{trial}}$.

Figure 1: Constant prediction (fixed) = 5, input mean $\in [0, 10]$, input standard deviation 0.
Figure 2: Inputs drawn from a uniform distribution. B and C: input standard deviation fixed at 4.5 when mean is varied, input mean fixed at 4.5 when input variance is varied.
Figure 2 Supplementary Fig. 1: Inputs drawn from different distributions with mean of 5 and variance of 4. Number of repetitions with different seeds: 20.
Figure 2 Supplementary Fig. 2: Inputs drawn from a uniform distribution with mean of 5 and variance of 4. Time step was 0.1. The connections from the PE neurons to the M or V neuron were altered by a factor γ drawn from a normal distribution. If not stated otherwise, the mean of this normal distribution was 1 and the variance 0, while the connection probability was 1. Number of repetitions with different seeds: 10.
Figure 2 Supplementary Fig. 3: Number of time steps were 4000. Number of identical mean-field networks: 1000.
Figure 3: C: $\sigma^2_{\mathrm{in}} = 0$ and $U(1, 9)$, D: $\sigma^2_{\mathrm{in}} = 5$ and $U(5, 5)$, F: $\sigma^2_{\mathrm{in}} = 0$ and $U(a, b)$ was parameterized such that the trail variability was 3.
Figure 3 Supplementary Fig. 1: Switch of input statistics occurs after 50 trails. State 1: stimulus $\in N(\mu_{\mathrm{in}}, 0)$ with $\mu_{\mathrm{in}} \in U(5, 5)$, State 2: stimulus $\in N(\mu_{\mathrm{in}}, 3)$ with $\mu_{\mathrm{in}} \in U(5, 5)$, State 3: stimulus $\in N(\mu_{\mathrm{in}}, 3)$ with $\mu_{\mathrm{in}} \in U(0, 10)$, State 4: stimulus $\in N(\mu_{\mathrm{in}}, 0)$ with $\mu_{\mathrm{in}} \in U(0, 10)$.
Figure 3 Supplementary Fig. 2: $\mu_{\mathrm{in}} = 5$ and $\sigma^2_{\mathrm{trial}} \in \{0, 0.75, 1.5, 2.25, 3\}$, $\sigma^2_{\mathrm{in}} \in \{3, 2.25, 1.5, 0.75, 0\}$. A: scaling factors of $w_{\mathrm{M}\leftarrow\mathrm{PE}}$ were 0.3 and 7.
Figure 4: $\mu_{\mathrm{in}} = 5$. A, top: $\sigma^2_{\mathrm{in}} = 0$ and $\sigma^2_{\mathrm{trial}} = 1$. A, bottom: $\sigma^2_{\mathrm{in}} = 1$ and $\sigma^2_{\mathrm{trial}} = 0$. C: $\sigma^2_{\mathrm{in}} = 1$ and $\sigma^2_{\mathrm{trial}} = 1$. D: $\sigma^2_{\mathrm{in}} = 5$ and $\sigma^2_{\mathrm{trial}} = 0$. Additional input (perturbation) was either fixed at 0.5 (A, D) or systematically varied between -1 and 1 (C), and was *on* for the last 50% of the trials. To estimate changes in baseline and gain of nPE and pPE neurons, we fitted a linear function to the PE neuron activity for the input range $[0, 2.5]$.
Figure 4 Supplementary Fig. 1: All parameters are from Figure 4 A.
Figure 4 Supplementary Fig. 2: In the default setting baseline was 0 and gain of PE neurons was 1. The results (left) were computed for baselines $\in [0, 3]$, while the results (right) were computed for gains $\in [0.5, 1.5]$. The results are based on the Eqs. (22), (25), (27) and (30).

Figure 4 Supplementary Fig. 3: All parameters are from Figure 4 D.
Figure 5: A: $\sigma_{\text{in}} \in \{1, 7\}$, $\mu_{\text{in}} \in U(15, 25)$. B: $\sigma_{\text{in}} \in [0, 8]$, $\mu_{\text{in}} \in U(15, 25)$, and $\sigma_{\text{in}} = 5$, $\mu_{\text{in}} \in U(15, b)$ with $b \in [20, 48]$. C: $\sigma_{\text{in}} \in \{2, 5\}$, $\mu_{\text{in}} \in U(15, 15)$. D: $\sigma_{\text{in}} = 0$, $\mu_{\text{in}} \in U(15, 25)$ or $U(10, 30)$. E: $\sigma_{\text{in}} = 0$, $\mu_{\text{in}} \in U(15, 25)$, or $\sigma_{\text{in}} = 5$, $\mu_{\text{in}} \in U(15, 15)$, Time steps per trail increased from 5000 to 10^4.
Figure 5 Supplementary Fig. 1: We used two different uniform distributions $U(15, 25)$ and $U(25, 35)$, and introduced scalar variability so that σ_{in} is a linear function of μ_{in}. Specifically, we chose $\sigma_{\text{in}} = [\mu_{\text{in}} - 14]_+$.

B Supporting analyses

We first describe a simplified model and show that the M neuron represents the mean, while the V neuron represents the variance of the feedforward input. We then investigate the impact of the gain and baseline of PE neurons on estimating the mean and variance. Furthermore, we use the simplified model to discuss the effect of neuromodulators in our network. Finally, we reveal the connection between the sensory weight and the contraction bias.

B.1 Activity of M and V neuron in a simplified model

To show that the M neuron encodes the mean, while the V neuron encodes the variance of the feedforward input, we resume a toy model in which the activity of the nPE and pPE neuron is replaced by its ideal output

$$\begin{aligned} r_{\text{nE}} &= [r_{\text{M}} - s_{\text{FF}}]_+ \\ r_{\text{pE}} &= [s_{\text{FF}} - r_{\text{M}}]_+ \end{aligned} \tag{15}$$

with s_{FF} denoting the time-dependent feedforward input. The activity of the M neuron can then be described as

$$\tau_M \cdot \frac{dr_{\text{M}}}{dt} = r_{\text{pE}} - r_{\text{nE}} \tag{16}$$

If $r_{\text{M}} \geq s_{\text{FF}}$, we get

$$\tau_M \cdot \frac{dr_{\text{M}}}{dt} = -r_{\text{nE}} = -r_{\text{M}} + s_{\text{FF}}. \tag{17}$$

If $r_{\text{M}} \leq s_{\text{FF}}$, we also get

$$\tau_M \cdot \frac{dr_{\text{M}}}{dt} = r_{\text{pE}} = -r_{\text{M}} + s_{\text{FF}}.$$

Hence, the activity of r_{M} is given by

$$r_{\text{M}} = \frac{1}{\tau_M} \int_0^t e^{-(t-x)/\tau_M} \cdot s_{\text{FF}}(x)\ dx \tag{18}$$

for zero activity at time $t = 0$. In the limit of $t \to \infty$ (steady state), this is exactly the exponential average of the feedforward input, $E(s_{\text{FF}})$.

With the simplified activity of the nPE and pPE neuron, the activity of the V neuron can then be described as

$$\tau_V \cdot \frac{dr_{\text{V}}}{dt} = -r_{\text{V}} + (r_{\text{pE}} + r_{\text{nE}})^2 = -r_{\text{V}} + (r_{\text{M}} - s_{\text{FF}})^2, \tag{19}$$

leading to the time-dependent solution

$$r_{\text{V}} = \frac{1}{\tau_V} \int_0^t e^{-(t-x)/\tau_V} \cdot [r_{\text{M}}(x) - s_{\text{FF}}(x)]^2\ dx. \tag{20}$$

In the limit of $t \to \infty$, r_{V} approaches $E(s_{\text{FF}} - E(s_{\text{FF}}))$.

B.2 Impact of PE neurons' gain on estimating mean and variance

The gains of the PE neurons, if not equal between nPE and pPE neuron on average, can bias the activity of both the M and V neuron. To show this, we resume the toy model from section B.1.

$$\begin{aligned} g_{\mathrm{pPE}}\ \langle r_{\mathrm{nPE}}\rangle &= g_{\mathrm{nPE}}\ \langle r_{\mathrm{pPE}}\rangle \\ g_{\mathrm{pPE}}\ \langle [s_{\mathrm{FF}}-P]_+\rangle &= g_{\mathrm{nPE}}\ \langle [P-s_{\mathrm{FF}}]_+\rangle \\ g_{\mathrm{pPE}} \int_P^\infty (x-P)\ f(x)\ dx &= g_{\mathrm{nPE}} \int_{-\infty}^P (P-x)\ f(x)\ dx. \end{aligned} \tag{21}$$

Here, P denotes the prediction encoded in the M neuron, and $f(x)$ is the distribution of feedforward inputs. In case of a uniform distribution, $f(x) = 1/(b-a)$ for $x \in [a,b]$ and 0 otherwise, we get

$$P = \begin{cases} \frac{a+b}{2} & \text{if } g_{\mathrm{nPE}} = g_{\mathrm{pPE}} = g \\ \frac{g_{\mathrm{pPE}}\cdot b - g_{\mathrm{nPE}}\cdot a + \sqrt{g_{\mathrm{nPE}}\ g_{\mathrm{pPE}}}\ (a-b)}{g_{\mathrm{pPE}} - g_{\mathrm{nPE}}} & \text{otherwise.} \end{cases} \tag{22}$$

Hence, the mean of the feedforward input is overpredicted when $g_{\mathrm{nPE}} < g_{\mathrm{pPE}}$. Similarly, the mean of the feedforward input is underpredicted when $g_{\mathrm{nPE}} > g_{\mathrm{pPE}}$ (Fig. S7).

Likewise, the variance is affected by the gain of the nPE and pPE neuron,

$$\begin{aligned} V &= \langle (r_{\mathrm{pPE}} + r_{\mathrm{nPE}})^2\rangle \overset{(i)}{=} \langle r_{\mathrm{pPE}}^2\rangle + \langle r_{\mathrm{nPE}}^2\rangle \\ &= g_{\mathrm{pPE}}^2\ \langle [s_{\mathrm{FF}}-P]_+^2\rangle + g_{\mathrm{nPE}}^2\ \langle [P-s_{\mathrm{FF}}]_+^2\rangle, \end{aligned} \tag{23}$$

where we assume (i) that both the nPE and pPE neuron have a zero baseline activity. In case of a uniform distribution, we get

$$\begin{aligned} V &= \frac{g_{\mathrm{pPE}}^2}{b-a} \int_P^b (x-P)^2\ dx + \frac{g_{\mathrm{nPE}}^2}{b-a} \int_a^P (P-x)^2\ dx \\ &= \frac{g_{\mathrm{pPE}}^2}{3}\cdot\frac{(b-P)^3}{b-a} + \frac{g_{\mathrm{nPE}}^2}{3}\cdot\frac{(P-a)^3}{b-a}. \end{aligned} \tag{24}$$

Inserting eqs. (22) yields

$$V = \begin{cases} \frac{(b-a)^2}{12} & \text{if } g_{\mathrm{nPE}} = g_{\mathrm{pPE}} = 1 \\ \frac{(b-a)^2}{3\ (g_{\mathrm{pPE}}-g_{\mathrm{nPE}})^3}\cdot\left[g_{\mathrm{nPE}}^2\cdot(g_{\mathrm{pPE}}-\gamma)^3 - g_{\mathrm{pPE}}^2\cdot(g_{\mathrm{nPE}}-\gamma)^3\right] & \text{otherwise.} \end{cases} \tag{25}$$

with $\gamma = \sqrt{g_{\mathrm{nPE}}\ g_{\mathrm{pPE}}}$. Hence, the variance of the feedforward input is overpredicted when $g_{\mathrm{nPE}} > 1$ or $g_{\mathrm{pPE}} > 1$. Similarly, the variance of the feedforward input is underpredicted when $g_{\mathrm{nPE}} < 1$ or $g_{\mathrm{pPE}} < 1$ (Fig. S7).

B.3 Impact of PE neurons' baseline on estimating mean and variance

The baselines of the PE neurons, if not equal between nPE and pPE neuron on average, can bias the activity of both the M and V neuron. By means of the toy model from section B.1, we can write

$$\begin{aligned} \langle r_{\mathrm{pPE}}\rangle &= \langle r_{\mathrm{nPE}}\rangle \\ \langle [s_{\mathrm{FF}}-P]_+ + p_0\rangle &= \langle [P-s_{\mathrm{FF}}]_+ + n_0\rangle \\ \int_P^\infty (x-P)\ f(x)\ dx + p_0 \underbrace{\int_a^b f(x)\ dx}_{=1} &= \int_{-\infty}^P (P-x)\ f(x)\ dx + n_0 \underbrace{\int_a^b f(x)\ dx}_{=1}. \end{aligned} \tag{26}$$

n_0 and p_0 denote the baseline activity of the nPE and pPE neuron, respectively. In case of a uniform distribution (c.f. B.2), we get

$$P = \frac{b+a}{2} + \frac{p_0 - n_0}{b-a}. \tag{27}$$

Thus, the M neuron encodes the true mean of the feedforward input only if $p_0 = n_0$. Hence, the mean is overpredicted if $p_0 > n_0$. Likewise, the mean is underpredicted if $p_0 < n_0$ (see Fig. S7).

With non-zero baseline activities, the steady state activity of the V neuron is given by

$$\begin{aligned} V &= \langle (r_{\mathrm{pPE}} + r_{\mathrm{nPE}})^2 \rangle \\ &= \langle [s_{\mathrm{FF}} - P]_+^2 \rangle + \langle [P - s_{\mathrm{FF}}]_+^2 \rangle + (p_0 + n_0)^2 + 2\ (p_0 + n_0)\ \left(\langle [s_{\mathrm{FF}} - P]_+ \rangle + \langle [P - s_{\mathrm{FF}}]_+ \rangle \right) \end{aligned} \tag{28}$$

In case of a uniform distribution $U(a, b)$, this expression yields

$$V = \frac{1}{3\ (b-a)} \left[(b-P)^3 + (P-a)^3 \right] + (p_0 + n_0)^2 + \frac{(p_0 + n_0)}{b-a} \left[(b-P)^2 + (a-P)^2 \right]. \tag{29}$$

Inserting the expression for P (Eq. 27) which is itself a function of the baseline activities, gives

$$V = \frac{(b-a)^2}{12} + \frac{(p_0 - n_0)^2}{(b-a)^2} \left(1 + 2\ \frac{p_0 + n_0}{b-a} \right) + (p_0 + n_0) \left(p_0 + n_0 + \frac{b-a}{2} \right). \tag{30}$$

Thus, for the V neuron to encode the variance unbiased, $n_0 = p_0 = 0$. The variance is overpredicted if either $n_0 > 0$ or $p_0 > 0$ (see Fig. S7).

B.4 Modelling the impact of neuromodulators on the sensory weight

We modeled the presence of a neuromodulator by simulating an additive excitatory input onto (groups of) interneurons. These interneurons, in turn, modulate the gain and baseline of PE neurons. As shown in sections B.2 and B.3, changes in the input-output transfer function of the PE neurons may bias the variance estimation in the network, and, hence, the sensory weight. Thus, understanding changes in the sensory weight requires an understanding of whether and how different types of interneurons change the PE neurons.

If a neuromodulator only acts on interneurons of the lower-level subnetwork, the sensory weight changes as a consequence of the modulated firing rates of the lower-level and higher-level V neurons. The lower-level V neuron is directly affected by the changes in the lower-level PE neurons and indirectly affected by changes in the M neuron of the same network. The higher-level V neuron is also affected by a neuromodulator acting in the lower-level subnetwork because the lower-level M neuron projects onto the neurons in the higher-level PE circuit. Hence, if the lower-level M neuron represents a biased mean $\mu \pm \delta\mu$, the variance estimation will be biased as well. This can be seen directly from the definition of the variance,

$$\begin{aligned} V &= \frac{1}{n} \sum_i \left(x_i - (\mu \pm \delta\mu) \right)^2 \\ &= \frac{1}{n} \sum_i \{ (x_i - \mu)^2 + \delta\mu^2 \mp 2\, \delta\mu\, (x_i - \mu) \} \\ &= V_{\mathrm{unmod}} + \delta\mu^2 \mp 2\ \delta\mu \left(\frac{1}{n} \sum_i x_i - \mu \right) \\ &= V_{\mathrm{unmod}} + \delta\mu^2 \end{aligned}$$

In contrast, if a neuromodulator only acts on interneurons of the higher-level subnetwork, the sensory weight changes as a consequence of the modulated firing rates of the higher-level V neuron. The higher-level V neuron is directly affected by the changes in the higher-level PE neurons and indirectly affected by changes in the M neuron of the same network.

Together, this suggests that whether the sensory weight decreases, increases, or remains the same in the presence of a neuromodulator depends on several factors:

- Does the neuromodulator act on the lower-level or higher-level subnetwork (that is, local impact), or does the neuromodulator act on both to the same degree (that is, global impact)?
- Which interneuron type/s is/are affected by the neuromodulator? And are these interneurons inhibited or excited by the neuromodulator?
- How are these interneurons embedded in the network, that is, what are the connectivity and the inputs to those neurons?

Hence, different neuromodulators may have the same effect on the sensory weight or the same neuromodulator may have different effects depending on brain area, species, etc..

B.5 Sensory weight and contraction bias

In the simulations, we define the bias as the trial-averaged difference between the weighted output and the true stimulus. For the sake of simplicity, we use r_{out} at the end of a trial as a proxy for the trial average in the subsequent analysis. Hence,

$$\text{bias} = r_{\text{out}}(T) - s. \tag{31}$$

To investigate how the bias depends on the sensory weight and potentially other factors, let us resume a toy model in which we assume that the prediction decays exponentially with time constant τ to a presented constant stimulus value, s,

$$P = P_0 \cdot \mathrm{e}^{-t/\tau_M} + s \cdot \left(1 - \mathrm{e}^{-t/\tau}\right) \tag{32}$$

with P_0 describing the initial value at time $t = 0$. Let us further assume that within a trial with trial duration T, the stimulus value changes n times ($T = n \cdot \Delta t$). The prediction during the presentation of the nth stimulus value can be expressed as

$$P_{\text{n}} = P_0 \cdot \mathrm{e}^{-\Delta t/\tau_M} + \left(1 - \mathrm{e}^{-\Delta t/\tau_M}\right) \sum_{i=1}^{n} s_i \cdot \mathrm{e}^{-(n-i)\cdot \Delta t/\tau_M} \tag{33}$$

To obtain an estimate for the prediction at the end of a trial, P_{n} must be averaged over the stimulus distribution, $\langle P_{\text{n}} \rangle_s$. For the sake of simplicity, let us assume the stimulus values are drawn from a uniform distribution $U\left(s - \frac{\sigma_{\text{S}}}{12}, s + \frac{\sigma_{\text{S}}}{12}\right)$. Moreover, we assume that the initial state, P_0, at the beginning of a new trial is drawn from a uniform distribution $U\left(\mu - \frac{\sigma_{\text{P}}}{12}, \mu + \frac{\sigma_{\text{P}}}{12}\right)$. With these assumptions, $\langle P_{\text{n}} \rangle_s$ is given by

$$\langle P_{\text{n}} \rangle_s = \mathrm{e}^{-\Delta t/\tau_M} \int_{\mu - \frac{\sigma_{\text{P}}}{12}}^{\mu + \frac{\sigma_{\text{P}}}{12}} P_0 \; f(P_0) \; dP_0 + \left(1 - \mathrm{e}^{-\Delta t/\tau_M}\right) \sum_{i=1}^{n} \mathrm{e}^{-(n-i)\cdot \Delta t/\tau_M} \int_{s - \frac{\sigma_{\text{S}}}{12}}^{s + \frac{\sigma_{\text{S}}}{12}} x \; f(x) \; dx \tag{34}$$

Solving the integrals yield

$$\langle P_{\text{n}} \rangle_s = \mu \cdot \mathrm{e}^{-T/\tau_M} + \left(1 - \mathrm{e}^{-\Delta t/\tau_M}\right) \sum_{i=1}^{n} \mathrm{e}^{-(n-i)\cdot \Delta t/\tau_M} \cdot s \tag{35}$$

Making use of the geometric series, the expression simplifies to

$$\langle P_{\text{n}} \rangle_s = \mu \cdot \mathrm{e}^{-T/\tau_M} + \left(1 - \mathrm{e}^{-T/\tau_M}\right) \cdot s$$

Inserting the expression in the equation for the weighted output yields

$$r_{\text{out}} = \left[\alpha_{\text{S}} \; \mathrm{e}^{-T/\tau_M} + \left(1 - \mathrm{e}^{-T/\tau_M}\right)\right] \cdot s + (1 - \alpha_{\text{S}}) \; \mathrm{e}^{-T/\tau_M} \; \mu$$

Hence, the bias in our toy model can be expressed by

$$\text{bias} = (1 - \alpha_{\text{S}}) \cdot \mathrm{e}^{-T/\tau_M} \cdot (\mu - s)\,.$$

The absolute slope $(1 - \alpha_{\text{S}}) \cdot \mathrm{e}^{-T/\tau_M}$ indicates how strong the bias is. It depends on the sensory weight α_{S}, the trial duration T and time constant τ. Please note that the sensory weight is a function of the trial duration itself (see Fig. 3F). However, for illustration purposes, we take α_{S} to be constant.

In this toy model, if the variance of the prediction is zero (that is, in a prediction-driven input regime), $\alpha_{\text{S}} \approx 0$, and, hence, the bias is $\mathrm{e}^{-T/\tau_M} \cdot (\mu - s)$. Thus, the bias is independent of the stimulus variance (see Fig. 5 D).

Likewise, if the variance of the sensory stimulus is zero (that is, in a stimulus-driven input regime), $\alpha_{\text{S}} \approx 1$, and, hence, the bias approaches 0 if the neurons reach their steady state. Thus, decreasing or increasing the trail variance does not have an effect on the bias (see Fig. 5 C).

C Supplementary Figures

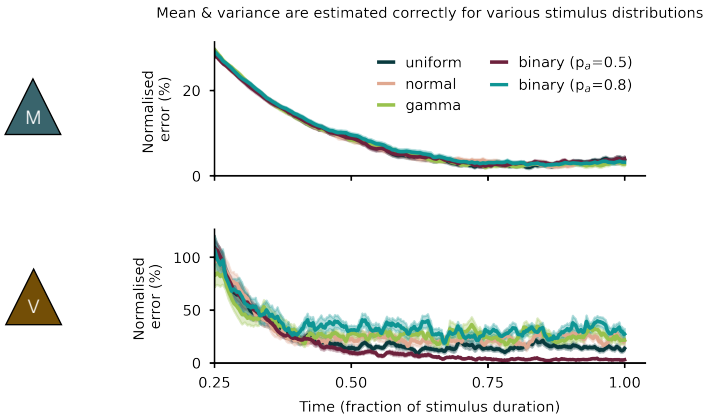

Figure S1. Estimating mean and variance of different stimulus distributions.
Top: The normalised absolute difference between the averaged mean and the activity of the M neuron decreases to a near-zero level for all stimulus distributions tested. Bottom: The normalised absolute difference between the averaged variance and the activity of the V neuron decreases with small differences between the distributions tested. Parametrisation of the uniform distribution as in Fig. 2.

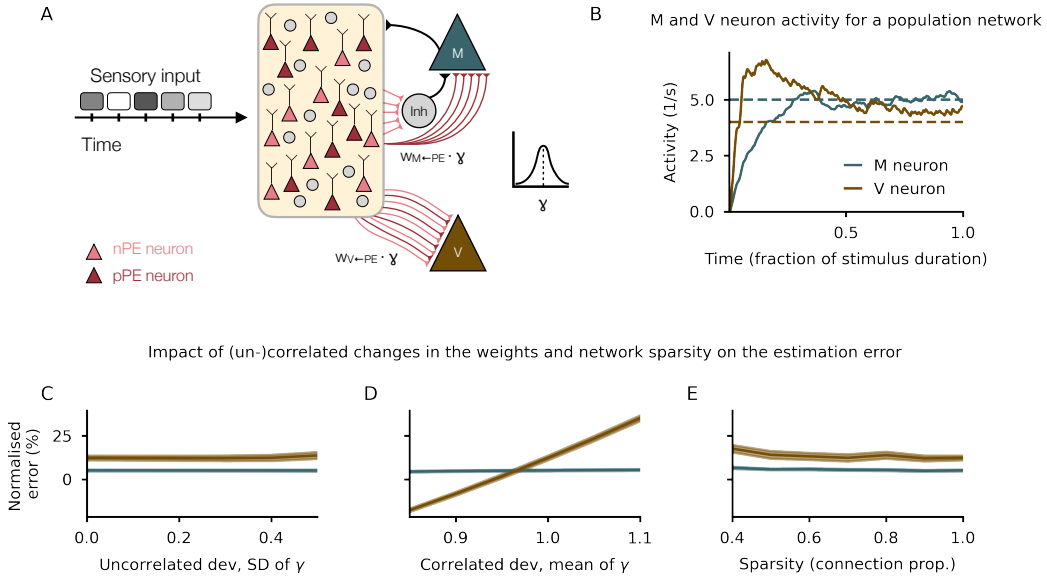

Figure S2. Estimating mean and variance of sensory stimuli in a rate-based population network.
(A) Illustration of the rate-based population network and the stimuli over time. **(B)** M and V neuron activities over time for one example parameterisation. **(C)** The normalised absolute difference between the averaged mean and the activity of the M neuron (dark green) or between the averaged variance and the activity of the V neuron (brown) for uncorrelated deviations, that is, increasing SD of γ (left), correlated deviations, that is, increasing mean of γ (middle), and the network sparsity.

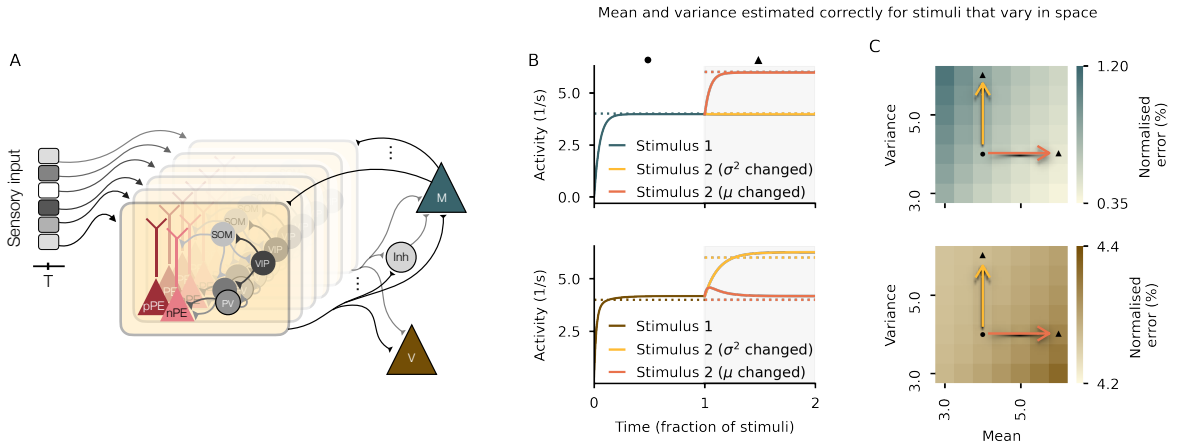

Figure S3. Estimating mean and variance of spatial stimuli.
(A) Illustration of a network estimating the mean and variance of a stimulus that varies across space. To simulate selectivity, the network comprises 1000 identical, uncoupled mean-field networks each receiving a different input value drawn from a uniform distribution. **(B)** Activity of M neuron (top) and V neuron (bottom) for 2 stimuli. The second stimulus does either differ in the mean (orange) or the variance (yellow) from the first stimulus (indicated in C). **(C)** The normalised absolute difference between the averaged mean and the activity of the M neuron (dark green, top) or between the averaged variance and the activity of the V neuron (brown, bottom) for a range of different stimulus statistics. The examples from B are shown with colored arrows and markers.

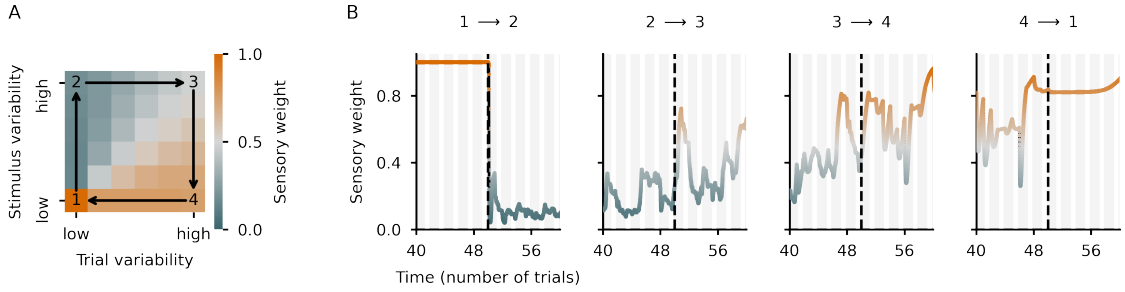

Figure S4. Dynamic variance estimation allows flexible adaptation to changes in the stimulus statistics and environment.
(A) Sensory weight for different input statistics. Numbers denote specific example statistics. Arrows denote the transitions between those statistics. **(B)** The sensory weight over time is shown for all transitions in (A). For the sake of clarity, we only show the trials 40 -60. The switch to new input statistics occurs at trial 50. Parameters are listed in the Supporting Information.

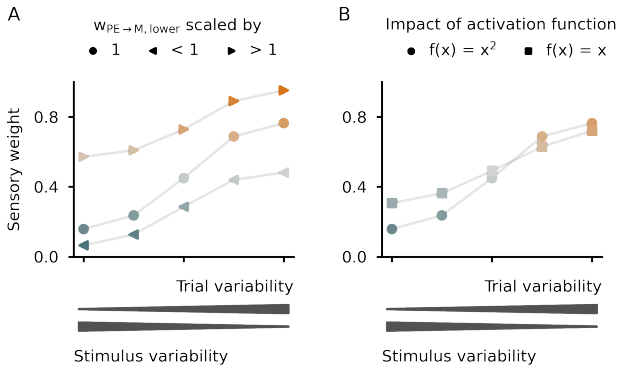

Figure S5. Perturbing the weighting of sensory inputs and predictions by altering network properties.
(A) The weights from the PE neurons to the M neuron in the lower-order subnetwork are scaled by a factor the 0.3 or 7, leading to a distorted sensory weight. If the update of the M neuron in the lower subnetwork is too slow (◀), the prediction is overrated. If the update of the M neuron in the lower subnetwork is too fast (▶), the sensory input is overrated. **(B)** The precise activation function for the V neurons does not have a major impact on the sensory weight. Only for inputs with high stimulus variability, the sensory stimulus is slightly overrated when the quadratic activation function is replaced by a linear, rectified activation function.

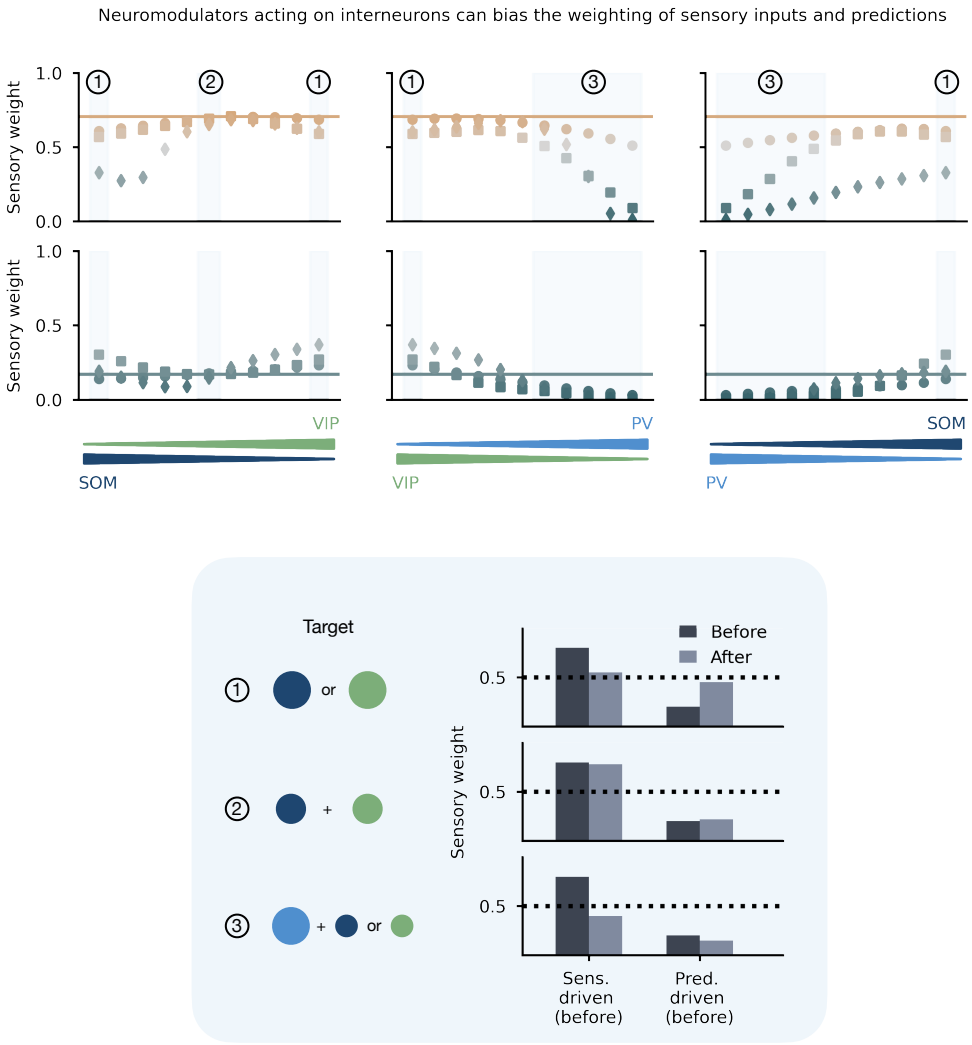

Figure S6. The impact of neuromodulators acting globally on groups of interneurons.
The sensory weight changes when groups of interneurons are targeted by a neuromodulator. Whether the sensory weight decreases or increases not also depends on the modulation strength (see Fig. 4) but also on how strongly which interneuron is targeted. As shown in Fig. 4, the sensory weight is pushed toward 0.5 if the VIP neuron is stimulated. The sensory weight generally decreases when PV neurons are the main target. Considered are two limit cases (upper row: more sensory-driven before modulation, lower row: more prediction-driven before modulation). The results are shown for three mean-field networks (see 4).

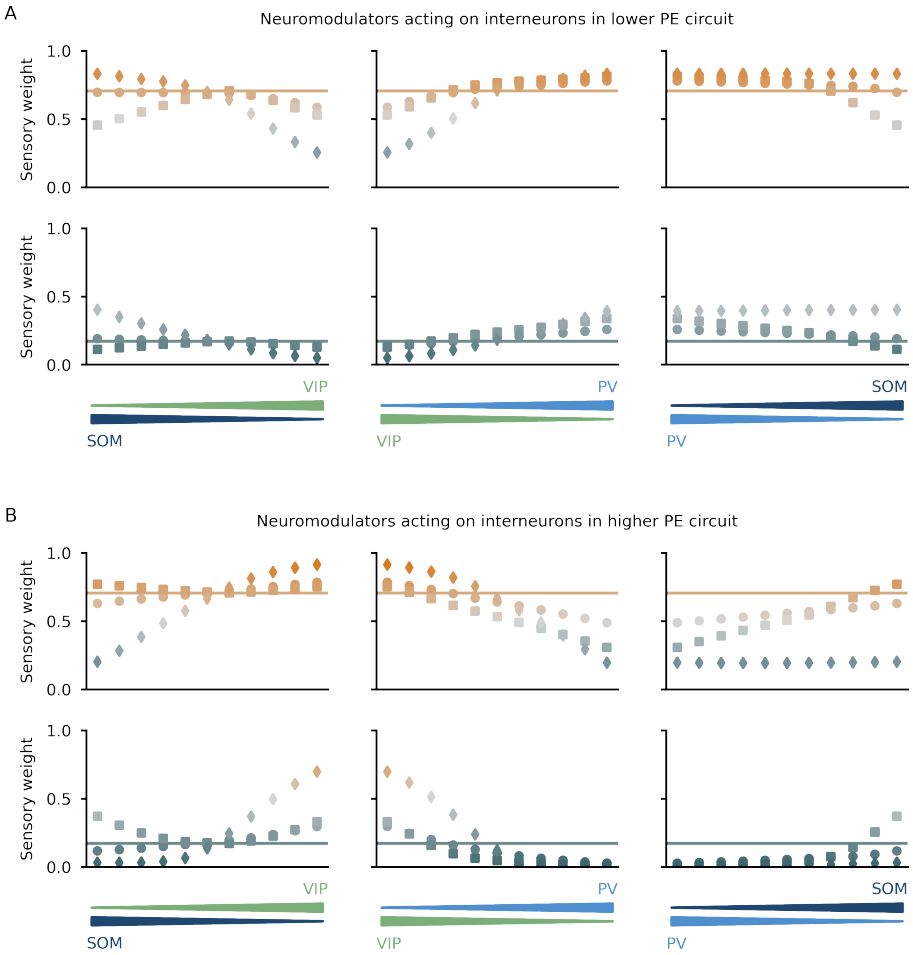

Figure S7. The impact of neuromodulators acting locally on groups of interneurons. **(A)** Sensory weight changes with neuromodulators acting on interneurons in the lower PE circuit. **(B)** Sensory weight changes with neuromodulators acting on interneurons in the higher PE circuit. In general, the changes in the sensory weight is the opposite of the changes seen for neuromodulators acting on the lower-level PE neurons. Simulation parameters, labels and colors as in Fig. 4.

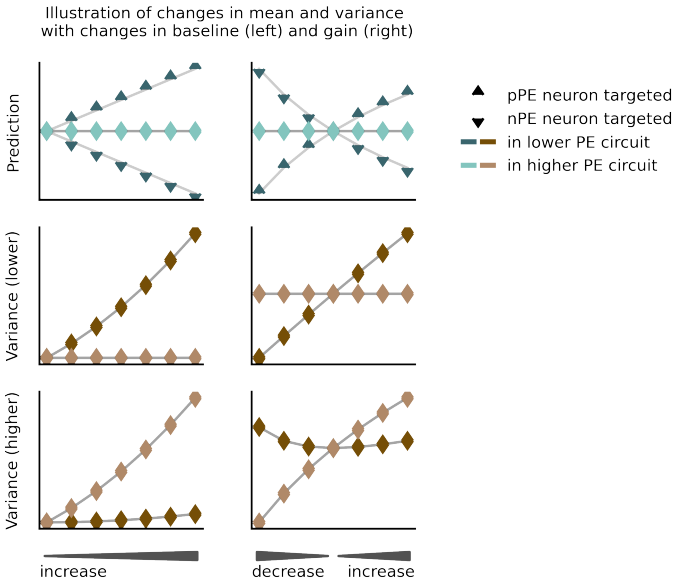

Figure S8. Biased mean and variance estimation by changing the baseline and the gain of nPE and pPE. In a toy model, described in sections B.2 and B.3, the contribution of gain and baseline to the changes in the mean and variance estimation are summarized. The results are based on the Eqs. (22), (25), (27) and (30).

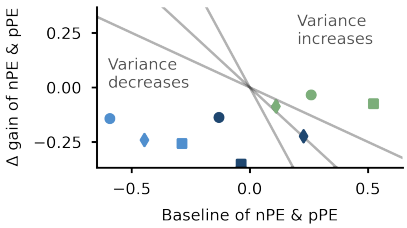

Figure S9. The combined changes in baseline and gain of all PE neurons determine the shift in the snesory weight.
Whether and how a neuromodulator changes the sensory weight depends on the interneuron targeted and the effect this interneuron has on the baseline and gain of both PE neurons, which in turn does depend on the network it is embedded in. For small inputs, changes in the baseline dominate, while for large inputs, the changes in the gains dominate the shift in the sensory weight.

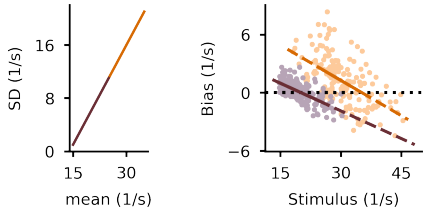

Figure S10. Including scalar variability in the model
When scalar variability is included, that is, the stimulus standard deviation depends linearly on the stimulus mean, the bias is larger for stimuli drawn from the upper end of the stimulus distribution than from the lower end.