

# Integrating feedforward and feedback inputs with prediction-error neurons

Loreen Hertäg<sup>1,\*</sup>, Katharina A. Wilmes<sup>2</sup>, Claudia Clopath<sup>3</sup>

1 Modelling 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](mailto: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. Furthermore, we investigate the causes of 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 context-dependency must also be reflected in the information processing of neural networks underlying intelligent behavior. For instance, when you walk down some staircase in your fully lit basement, your brain might entirely rely on the feedforward (bottom-up) input your senses receive (Fig. 1, 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. 1, middle).

The importance of these feedback inputs has been emphasized by observations showing that top-down projections outnumber feedforward connections (XXX) and that they modulate (XXX) or even entirely drive (XXX) neuron activity. But how do neural networks switch between a feedforward-dominated and a feedback-dominated processing mode? And how do neural networks 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. 1, 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 (XXX). 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 xxx (XXX). It is conceivable that the same concepts can be employed for the weighting of sensory inputs and predictions thereof (XXX). A central point in the weighting of inputs is the estimation of 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 hypothesised 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. When those predictions are wrong, the resulting PEs allow the network to revise the model of the world, thereby ensuring that the predictions are more accurate (XXX). Experimental evidence suggests that these PEs may be represented in the activity of distinct groups of neurons, termed PE neurons (XXX). Moreover, these neurons may come

in two types when excitatory neurons exhibit near-zero, spontaneous firing rates (XXX). Negative PE (nPE) neurons mainly increase their activity when the prediction is *stronger* than the sensory input, while positive PE (pPE) neurons mainly increase their activity when the prediction is *weaker* than the sensory input. Indeed, it has been shown that excitatory neurons in layer 2/3 of rodent primary sensory areas can encode negative or positive PEs (XXX).

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. In line with multisensory integration, predictions are weighted more strongly than the sensory stimuli when the environment is stable (that is, predictable) but the sensory inputs are noisy. Moreover, we show that predictions are integrated more strongly after a change in the environment, even 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 like parvalbumin-expressing (PV), somatostatin-expressing (SOM), and vasoactive intestinal peptide-expressing (VIP) interneurons (XXX). In a computational model, these interneurons have been shown to establish a multi-pathway balance of excitation and inhibition that is the basis for nPE and pPE neurons (XXX). By breaking this balance, the excitatory neurons change their baseline firing rate and gain, leading to a biased variance estimation. Finally, we show that this weighting can be understood as the neural manifestation of the contraction bias (XXX).

## Results

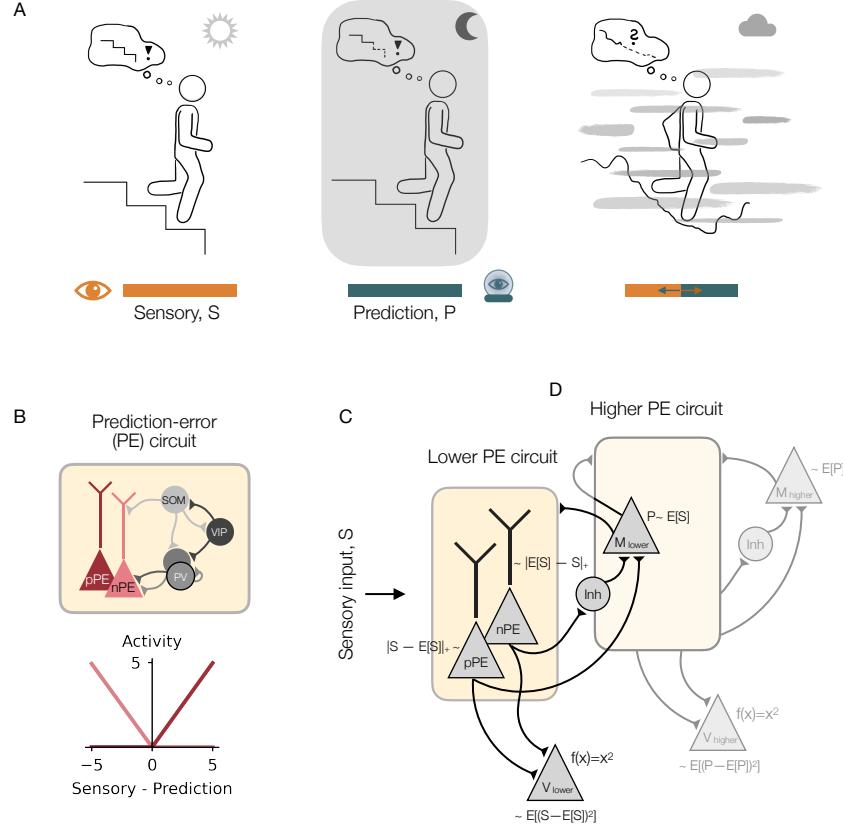
### nPE and pPE neurons as the basis for estimating mean and variance of sensory stimuli

We hypothesise that the distinct response patterns of negative and positive prediction-error (nPE/pPE) neurons represent the backbone for estimating the mean and the variance of sensory stimuli. nPE neurons only increase their activity relative to a baseline when the sensory input is weaker than predicted, while pPE neurons only increase their activity relative to a baseline when the sensory input is stronger than predicted. Moreover, both nPE and pPE neurons remain at their baseline activity when the sensory input is fully predicted (XXX). Assuming that 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.

To test our hypothesis, we studied a rate-based mean-field network the core of which 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 balanced because it has been shown that this E/I balance is necessary for nPE and pPE neurons to emerge (XXX, see Methods).

In addition to this core circuit, we model a memory (M) neuron that perfectly integrates the activity of the PE neurons (Fig. 1C). In accordance with XXX, 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). The M neuron connects to the apical dendrites of the PE neurons and some of the interneurons (here, VIP and PV neurons, see Methods for more details). In this network, the M neuron serves as a prediction that is dynamically updated when new sensory information is available. We furthermore simulate a downstream neuron (termed V neuron), modeled as a leaky integrator with a squared activation function, that receives excitatory output 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 mean and variance in the respective neurons, we stimulate it with a sequence of step-wise constant inputs drawn from a uniform distribution (Fig. 2A), assuming 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). This is in contrast to the three interneurons that strongly increase with stimulus mean while they only moderately increase with stimulus variance (Fig. 2C). The activity of the memory neuron M gradually approaches the mean of the sensory inputs (Fig. 2D, middle), while 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.

activity of the V neuron approaches the variance of the inputs (Fig. 2E, middle). This is true 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 larger input variances, while the estimated variance is fairly independent of the input statistics tested.

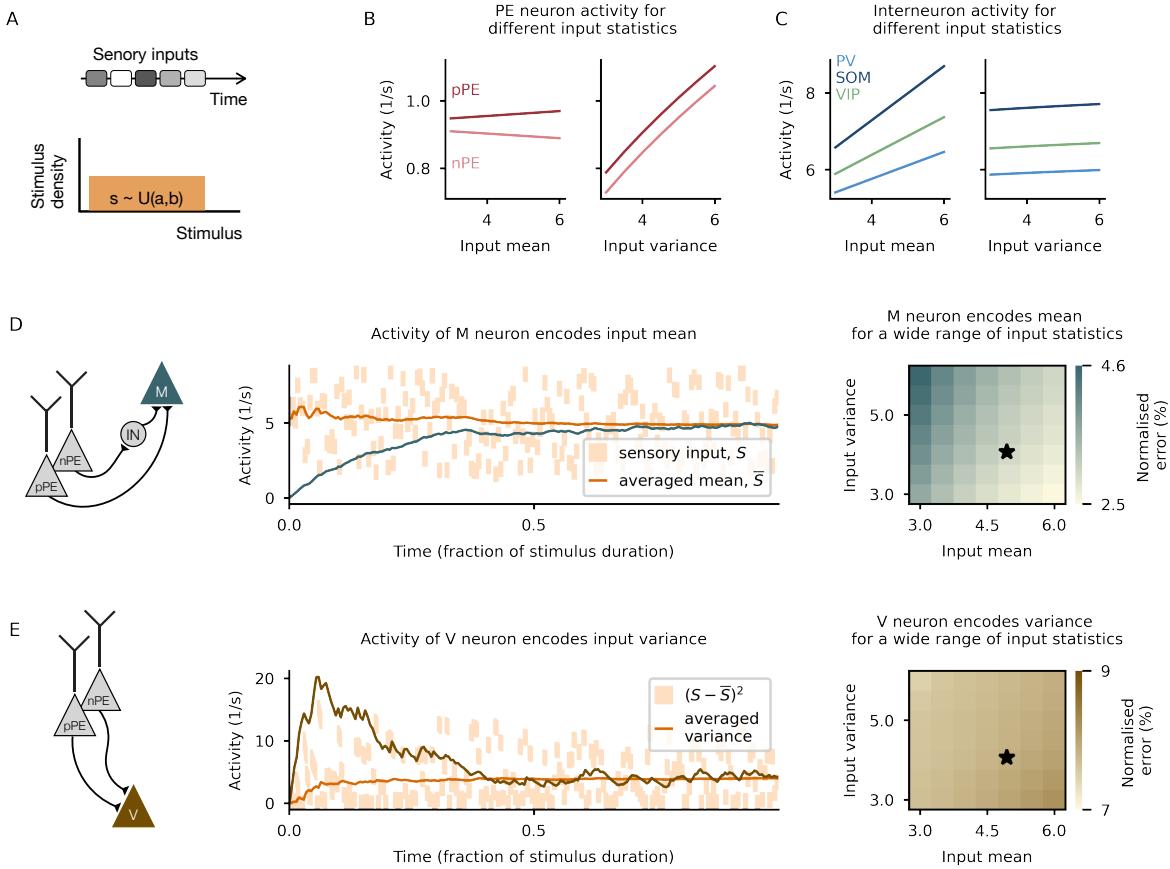
XXX coming soon: Paragraph on network beyond mean-field XXX

In summary, nPE and pPE neurons can be the basis to estimate the mean and the variance of sensory stimuli that vary over time.

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

Following the ideas of Bayesian multisensory integration (XXX), the weighting of sensory stimuli and predictions thereof would require knowledge of their variances. As we have shown in the previous section, the variance of the sensory stimulus can be estimated using PE neurons. We hypothesise that the same principles apply to computing the variance of the prediction. Hence, we augment the network with a *higher* PE circuit that receives output synapses from the M neuron of the *lower* PE circuit (Fig. 1D). Both subnetworks are modeled the same, except that the M neuron in the higher PE circuit evolves more slowly 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. In each trial one stimulus is shown to the network. To account for the stimulus



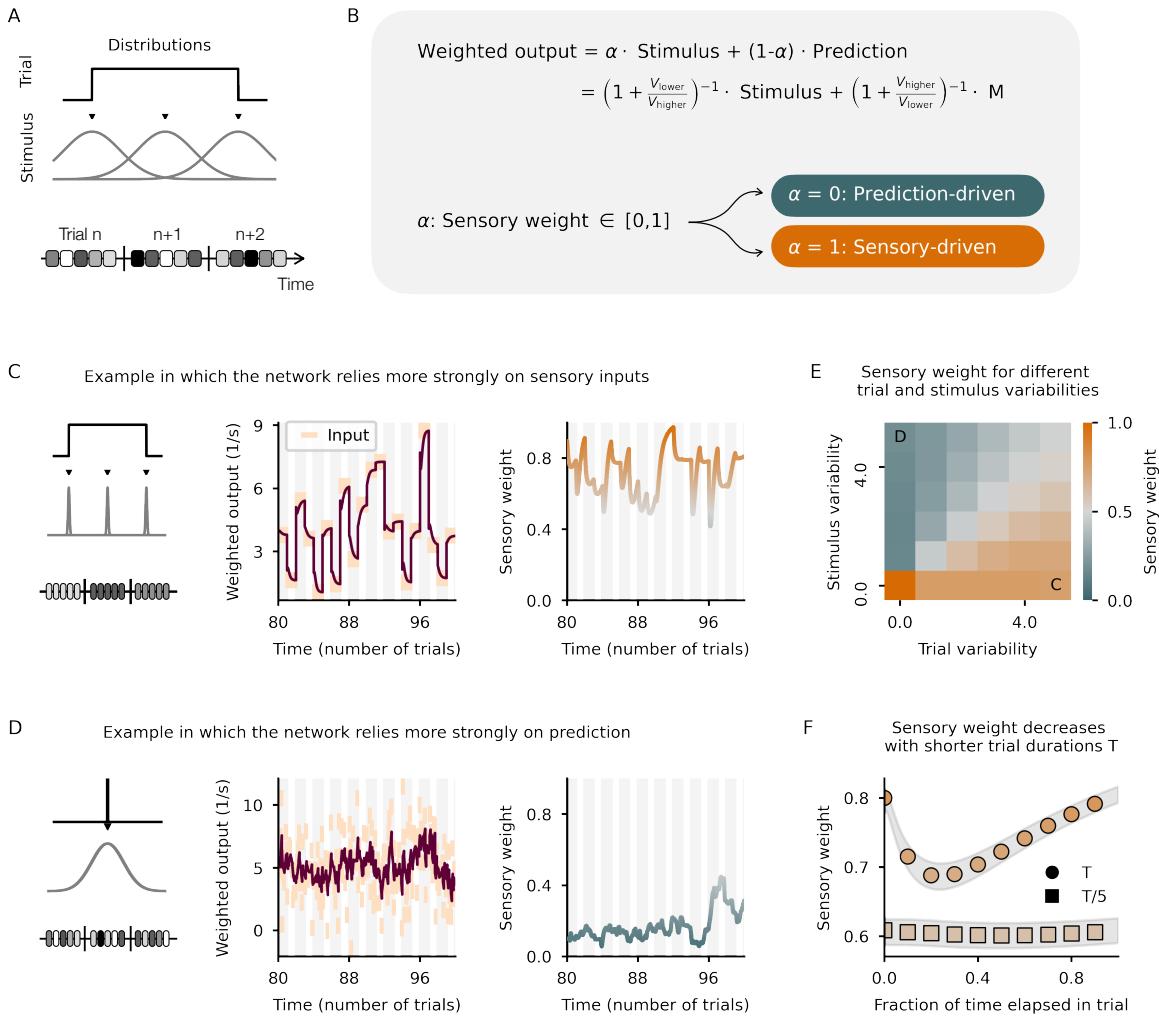
**Figure 2. Prediction-error neurons as the basis for estimating mean and variance of sensory stimuli.**

(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.

variance, each stimulus is composed of  $n$  constant values drawn from a normal distribution with mean  $\mu_{\text{stim}}$  and variance  $\sigma_{\text{stim}}^2$ , and presented one after the other. To account for potential changes in the environment, in each trial, we draw  $\mu_{\text{stim}}$  from a uniform distribution (Fig. 3A). Hence, the inputs change on two different time scales, with stimulus variability (faster time scale) and trial variability (slower time scale).

As expected, the neurons' activity increase for both stimulus and trial variances (Fig. 3B). While the neurons in the lower PE circuit increase more strongly with stimulus variability, the neurons in the higher PE circuit increase more strongly with trial variability, indicating that the different subnetworks process different aspects of the inputs. We first consider two limit cases. In the first limit case, a different but low-variance stimulus is presented in each trial (Fig. 3C, left). In line with the ideas of multisensory integration (XXX), the network should therefore follow the sensory inputs closely and ignore the predictions. When we arithmetically calculate the weighted output (Fig. 3C, middle) based on the feedforward and feedback inputs, and the sensory weight (Fig. 3C, right), the network correctly represents mostly the sensory input (for more details, see Methods). In the second limit case, the same but 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).

In a next step, to validate the network responses more broadly, we systematically varied the trial



**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  $\alpha$  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.

and stimulus variability independently. If both variances are similar, the sensory weight approaches 0.5, reflecting equal contribution of sensory inputs and predictions 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. This is reversed if the stimulus variance is smaller than the trial variance (Fig. 3E). Because the network dynamically estimates the mean and variances of the sensory input and the prediction, the weighted output and the sensory weight changes accordingly when the input statistics changes (Fig. S4).

The first limit case (Fig. 3C) shows that even in a sensory-driven input regime, the prediction is weighted more at the beginning of a new trial than in the steady state. This is further confirmed in simulations in which the trial duration was shortened. For those simulations, the prediction even outweighs the sensory input, reflected in a very low sensory weight (Fig. 3F). This suggests that predictions influence

neural activity more significantly in experiments that rely on very fast stimulus changes.

It has been shown [speculated?], that sensory inputs or predictions are overrated in some psychiatric disorders (XXX). We thus wondered which network properties might bias the estimation of the variances, and, consequently, the weighting of different input streams. In our network, the M neuron evolves faster in the lower subnetwork than in the higher one. We identified the speed at which the M neurons are updated with 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 M neuron evolves too slowly, the prediction is overrated. In contrast, if the M neuron incorporates new information too quickly, the sensory input is overrated (Fig. S5A). While the speeds at which the activity of the M neurons evolve may underlie pathological weighting of inputs, the precise activation function of the V neurons is less pivotal. When we replaced the squared activation function with a linear, rectified function, the V neurons do not encode the variance but the averaged 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. The model shows that predictions are trusted more strongly at the beginning of a new stimulus, and if sensory inputs are noisy on a short time scale while predictable on longer time scales.

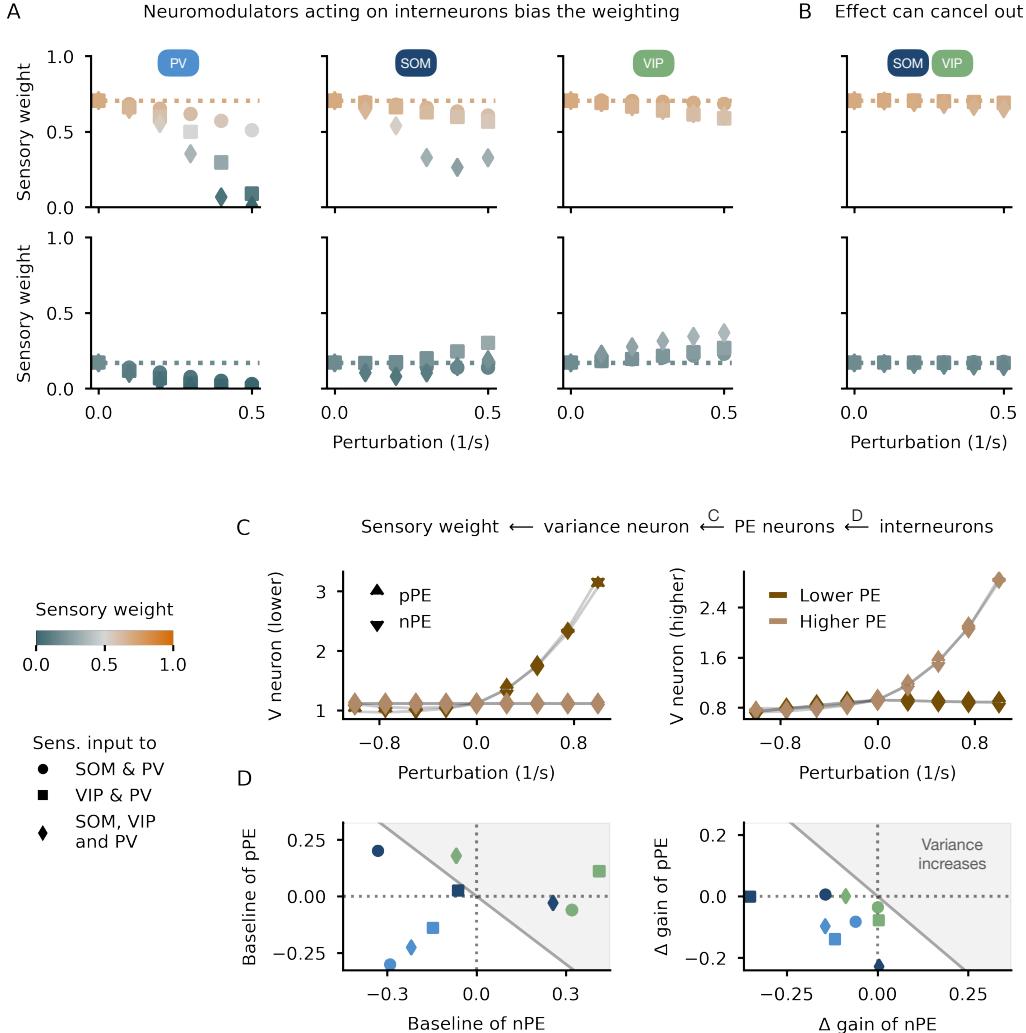
## Biasing the weighting of sensory inputs and predictions by neuromodulators

The brain's flexibility and adaptability are not least because a plethora of neuromodulators influence the activity of neurons in a variety of ways (XXX). A prominent target of neuromodulatory inputs is inhibitory neurons (Cardin 2019, XXX). Moreover, distinct interneuron types are differently (in-)activated by those neuromodulators. For instance, it has been shown that XXX (XXX). 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 neuromodulators by injecting an additional excitatory input into one or two interneuron types. 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 three different mean-field networks that differ with respect to 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 (XXX). Across the different mean-field networks tested, activating a SOM or VIP neuron individually forces the networks to weigh both inputs more equally. As a consequence, predictions are overrated in a sensory-driven input regime. Similarly, sensory inputs are overrated in a prediction-driven input regime. Interestingly, when both interneuron types are activated to the same degree, this effect disappears (Fig. 4A, left). In contrast, stimulating PV neurons biases the network's output towards predictions. This effect is even more pronounced when PV and SOM, or PV and VIP neurons are activated simultaneously (Fig. 4A, middle and right).

In the previous simulations, we assumed that a neuromodulator acts globally, that is, on the interneurons in both the lower and the higher PE circuit. While this agrees with experimental data showing that XXX (XXX), we note that neuromodulators may also act more locally. The effect of stimulating an interneuron type in the lower PE circuit on the sensory weight is mostly the opposite of activating the same interneuron in the higher PE circuit (Fig. S6). For instance, the sensory inputs are overrated when the higher-level VIP neuron is activated, while the prediction is overrated when the lower-level VIP neuron is activated. When VIP and SOM neurons are stimulated equally, the sensory weight remains unchanged, independently of which PE circuit is targeted by the neuromodulator.

What are the mechanisms that give rise to these effects? And how do the combined local changes give rise to the global one observed in our network simulations? The sensory weight is chosen to be a function of the V neurons of the lower and higher PE circuit. Hence, any changes to the sensory weight result from changes to the neurons encoding the variances (Fig. 4B). In our network, the V neurons only receive excitatory output 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. At first, this is counterintuitive because the V neuron in the higher subnetwork does not receive direct synapses from the PE neurons in the lower subnetwork. However,



**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.

the activity of the lower-level M neuron encoding the prediction increases with an excitatory input onto the pPE neuron and decreases with an excitatory input onto the nPE neuron (the opposite is true for an inhibitory input). Because neurons in the higher PE circuit receive synapses from the lower-level M neuron, the activity of the higher-level V neuron is also affected. In contrast, stimulating either PE neuron in the higher subnetwork increases the activity of the higher-level V neuron but leaves the lower-level M and V neurons unaffected (Fig. 4C).

Stimulating PE neurons may cause both an increase in the baseline activity and a change in the neuron's gain. To disentangle both effects, we illustrate each contribution separately using a mathematically tractable toy model (see Methods for more details). The variance estimated in the lower subnetwork increases with both increasing baseline and gain of the lower-level PE neurons. In contrast, when the gain of those PE neurons decreases, so does the variance. Similarly, the variance estimated in the higher subnetwork is equally influenced by changes in baseline and gain of the higher-level PE neurons. Moreover, changes to the baseline and the gain of the lower-level PE neurons increases the higher-level

variance as a result of a biased prediction. Furthermore, the mean of the sensory stimuli is overpredicted (underpredicted) when the baseline or the gain of the lower-level pPE (nPE) neuron increases, or when the gain of the lower-level nPE (pPE) neuron decreases. In summary, if an interneuron causes an additional inhibitory input to a PE neuron, the neuron's gain is reduced. If an interneuron causes an additional disinhibitory input to a PE neuron, the neuron's baseline and gain are increased (Fig. 4D).

This suggests that to understand the effect of neuromodulators on the sensory weight, we need to unravel the effect of interneuron activation on baseline and gain of PE neurons. To this end, we stimulated either PV, SOM, or VIP neurons independently for all three mean-field networks and measured the changes to baseline and gain of both PE neurons (Fig. 4E). In all three networks tested, activating PV neurons decreases both quantities, leading to a decrease in the estimated variance (Fig. 4E & Fig. S7). Stimulating SOM or VIP neurons decreases the overall gain but increases the baseline activity of the PE neurons. Whether and how much the gain of the nPE or pPE neuron is reduced depends on the inputs onto SOM and VIP neurons, and the connectivity they make with other neurons in the network. Similarly, how much the baseline is elevated depends on the specifics of the mean-field network (Fig. 4E & Fig. S7). Hence, whether stimulating the SOM or VIP neuron decreases or increases the activity of the V neuron 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.

### The contraction bias as a result of the weighted integration of sensory inputs and predictions

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 (Fig. 5A). This *bias towards the mean* has been reported in different species and modalities (XXX).

The weighted output of our network can be interpreted as a neural manifestation of the contraction bias (see Methods for a thorough analysis). The bias increases with stimulus variance (Fig. 5B), decreasing the slope of the linear fit modeling the relationship between the true and estimated stimuli (Fig. 5B, right; compare with Fig. 5A). In contrast, the bias decreases with trial variance, so that the slope of the linear fit approaches 1 (Fig. 5B right).

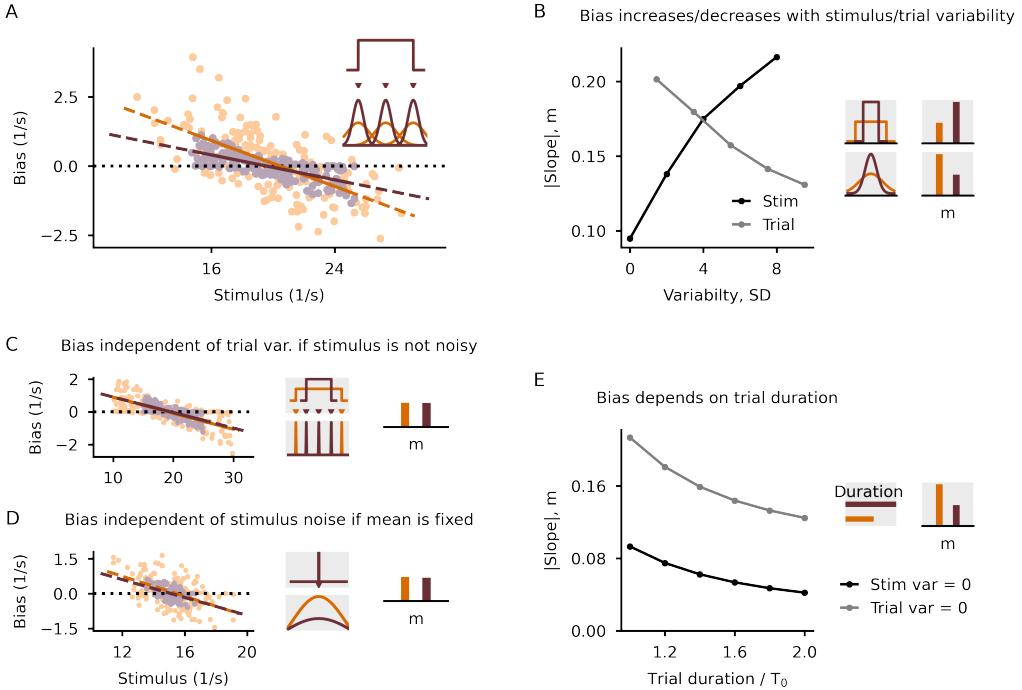
What are the underlying network factors that contribute to this phenomenon? To disentangle the potentially different sources of the bias, we first simulated a network without stimulus variability (variance set to zero) for two different trial variabilities. In this case, a contraction bias emerges but is independent of the volatility of the environment (Fig. 5C, left). We show mathematically that the bias results from the transient neuron activity before reaching a steady state, and vanishes if the trial duration approaches infinity (see Methods for more details, and Fig. 5D). We next resume the limit case in which the same but high-variance stimulus is shown in every trial. In this case, the weighted output exhibits a contraction bias that is largely independent of the stimulus variance (Fig. 5C, right and Fig. 5D). As shown mathematically (see Methods), the bias results from the finite trial duration and the tendency to weight the prediction more strongly than the sensory inputs.

So far, we assumed that the stimulus variance is independent of the trial mean. A consequence of this choice is that the bias on either end of the stimulus distribution is largely the same (but with reversed signs). However, behavioral (neural?) data (XXX) 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 trial mean. In these simulations, as expected, the bias increases for a stimulus distribution shifted to higher trial means (Fig. 5E).

In summary, the weighted integration of sensory inputs and predictions can be interpreted as a neural manifestation of the contraction bias. Both stimulus and trial variability contribute to the contraction bias but the underlying mechanisms differ.

## 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 ???. Conflicting information may be caused by noise in the sensory feedforward inputs or by changes in the environment



**Figure 5. Mechanisms underlying the contraction bias.**

(A) Illustration of the contraction bias. The estimated input is shifted towards the mean of the input distribution. Hence, a linear curve fitted to the data has a slope below 1 (left). The bias is biggest at the end of the stimulus distribution (right). (B) Contraction bias in the model. Left: Example for two different stimulus variabilities. Right: As a consequence of the sensory weight, the slope decreases with stimulus variability (bias increases) and increases with trial variability (bias decreases). Stimulus statistics: XXX. (C) The bias is independent of trial variability when the stimulus variability is zero (left). Equally, the bias is independent of the stimulus variability when the trial variability is zero (right). Stimulus statistics: XXX. (D) The slope depends on the trial duration. If the sensory weight is 1, the slope is independent of the trial duration and the bias vanishes (dashed-dotted line). If the sensory weight is 0, the slope depends on the trial duration and only reaches 1 if the trial duration approaches infinity. (E) To ensure a larger bias for stimuli drawn from the upper end of the stimulus distribution than from the lower end, scalar variability as observed experimentally is needed.

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.] [yon2021precision.

This idea is consistent with Bayesian theories on the optimal integration of multiple sensory cues (aka multisensory integration). ?? 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. XXX name other examples in human and other species XXX. These behavioral studies were accompanied by neural recordings showing that XX

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 ?? . Indeed, there is evidence that the brain can choose to rely more on top-down or bottom-up signals ?? . 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 ?? . XXX We suggested that xxx Here summarise your contributions XXX

XXX put this in here somewhere XXX 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. 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 ?? . 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 ???. In our model, those modulatory signals act through interneurons ?? 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 underestimates the uncertainty of the sensory input in a prediction-driven regime. Hence, the system leans toward weighting sensory inputs and predictions more equally. Increasing SOM neuron activity is comparable to modulating VIP neurons. However, the equalising effect is less robust and seems to depend on the network specifics (Fig. ??A). Interestingly, when SOM and VIP neuron activity is modulated to the same degree, the weighting remains unaffected, suggesting that the individual contributions cancel (Fig. ??B). We show that these findings can be best explained by changes in the baseline and gain of PE neurons as a result of the modulation of interneuron activity (Fig. ??D). 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 variance of the sensory cues and the predictions, as well as the trial duration. In our model, if the stimulus variance is zero (perfect sensory-driven network regime), the neural contraction bias is independent of the trial variability. In that case, the bias reflects the slow transition from the last stimulus to the new one, and, hence, would diminish if the trial is sufficiently long. On the other hand, if the trial variance is zero (perfect prediction-driven network regime), the neural contraction bias is independent of the stimulus variability and a function of the trial duration. However, we note that in our model, we only consider the 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 ??.

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 & Schultz, 2010, XXX). 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 & Pouget, 2004; Ma et al., 2006, 2008). The neurons' activities represent how close the sensory input is to the preferred (predicted) input of each neuron.

There has been evidence that indeed (population of) neurons can encode (un-)certainty. Neurons in the pulvinar (Komura et al., 2013), the parietal cortex (Kiani and Shadlen, 2009) or the frontal cortex (Kepecs et al., 2008, Middlebrooks and Sommer, 2012) encode the uncertainty of perceptual decisions ???. In addition, it has been shown that neurons in the orbitofrontal cortex of rats signal confidence in decisions ???. Moreover, neural correlates of (expected) uncertainty have been found in regions of the prefrontal cortex 43 ??, the rat orbitofrontal cortex [52,53??] or the dorsal striatum in monkeys78 ???. Interestingly, the accuracy of memory recalls is represented (?) in single neurons of the human parietal and temporal lobes (Rutishauser et al., 2015, 2018) ... in ??.

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 hypothesised in predictive processing theories ???. While it has been hypothesised that these internal representation neurons might be found in deeper L5 neurons ??, there is also evidence that a group of excitatory L2/3 neurons integrates over negative and positive prediction errors ??.

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 ???. Moreover, while first only hypothesized theoretically ??, the presence of two types of PE neurons, the negative and positive PE neurons, has been confirmed in a number of recent studies ???. In our model, the nPE neuron inhibits the memory neuron while the pPE neuron excites it, in line with ???. 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 (XXX see refs in Siegle 2021 Allen).

However, 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. ??). It has been hypothesized that symptoms in psychiatric diseases may be a result of an erroneous uncertainty estimation ???. For instance, hallucinations may be the result of an underestimation of the expectation uncertainty or an overestimation of the sensory uncertainty. Conversely, a fixation on a stable environment even when the sensory cues indicate a switch in the environment, may be the consequence of an overestimation of the

expectation uncertainty or an underestimation of the sensory uncertainty ??.

A popular hypothesis is that neuromodulators shape the weighting of sensory inputs and predictions thereof ?? . Theoretical work by ?? suggests that acetylcholine (ACh) correlates with expected uncertainty, while noradrenaline (NA) correlates with unexpected uncertainty. Expected uncertainty is usually considered to be related to 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 ?? . While in our network the stimulus and trial variability can only be loosely interpreted as 'expected' and 'unexpected' uncertainty, we want to XXX broader context and highlight differences XXX

It is assumed that NA increases in more volatile environments and enhances bottom-up processes [20,21 in ??, yon2021precision]. In line with this idea, NA blockade impairs cognitive flexibility [24,25??]. In recent work by ??, 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 ?? . A major target for noradrenergic inputs is SOM neurons whose activity increases in the presence of NA (reviewed in Urban-Ciecko 2016: Fanselow, 2008; Urban-Ciecko 2015; Chen 2015; Xiang 1998; reviewed in Hattori 2017: Paspalas 1999; reviewed in Swanson in Maffei 2019: Kawaguchi and Shindou, 1998). In our model, activating SOM neurons globally does not enhance 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. However, increasing SOM activity only in the lower-order circuit shows a slight enhancement of the sensory weight (Fig. ??).

Similarly to NA, ACh has also been shown to enhance bottom-up, feedforward inputs (Gil et al., 1997; Hasselmo et al., 1996; Hsieh et al., 2000; Kimura et al., 1999; Kobayashi, 2000, reviewed in Yu and Dayan 2005 or Marshall 2016). For instance, subjects relied more strongly on prior beliefs when given cholinergic receptor antagonists (reviewed in Bruckner 2022: Marshall 2016). A major target for cholinergic input are VIP neurons whose activity increases in the presence of ACh (reviewed in Wester 2014; reviewed in Hattori 2017: Fu 2016, Kuchibhotla 2017, Nelson 2016; reviewed in Swanson in Maffei 2019: Paul et al., 2017). 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 circuit does generally enhance sensory bottom-up inputs (Fig. ??).

Neuromodulatory systems have been suggested to gate plasticity ?? . In a recent study by ??, locus caeruleus (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 ?? . In our model, we do not consider precision-weighted PEs (but see ??). 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 of incorporating new sensory evidence when environments change (??).

As for any computational model, we brush over a number of biological details in order 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, evidence suggests that prediction-errors itself are precision-weighted ?? . xx Kelly Diederend ?? xxx show that neural signatures of reward-prediction errors are augmented when the outcome is reliable and attenuated when the outcome is uncertain. XXX Katha's work XXX

Furthermore, while we provide a neuronal circuit model for estimating the mean and the variance of both sensory signals and predictions, we do not model the weighting of inputs explicitly. A respective neural circuit model would require divisive inhibition. XXX (look in paper for neural implementation of Kalman filter) xxx have shown how this can be done with inhibitory interneurons.

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, ?? 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; Burr & Alais, 2004). ... in ??), but trust vestibular information more than visual information about head direction (Butler et al., 2010 ??), a finding also observed for monkeys (Angelaki 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. Moreover, it has been suggested that the brain may use different strategies to combine signals depending on the task demands ?? . 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.

A number of normative models have been proposed for state estimation and prediction under uncertainty ??, ranging from the classical Kalman filter to more recent models like *Bayes Factor Surprise* ???. For instance, Bayes factor surprise formalizes the trade-off between integrating new observations in an existing belief system and resetting this belief system with the new evidence. The surprise factor captures how much an animal's current belief deviates from the new observation.

In more recent years, normative model have been squared with biological constraints. For instance, in a paper by ??, 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.

XXX Come back to your model XXX highlight the nice thing about it and then find some concluding remarks

## Models and methods

### Network model

The mean-field network model consists of a *lower* and *higher* 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  $V$ .

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 ( $\underline{r}_{PE}^{low}$  and  $\underline{r}_{PE}^{high}$ ) are given by

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

with

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

We follow the notation that column and row vectors are indicated by small 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  $\underline{r}_{PE}^{loc} = [r_{nE}^{loc}, r_{pE}^{loc}, r_{nD}^{loc}, r_{pD}^{loc}, r_{PV_1}^{loc}, r_{PV_2}^{loc}, r_{SOM}^{loc}, r_{VIP}^{loc}]$  with loc  $\in [\text{low}, \text{high}]$  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  $\underline{I}_{PE}$ . The connection strengths between the pre-synaptic neuron (population) and the neurons of the PE circuit are denoted by  $W_{PE \leftarrow pre}$  and  $\underline{w}_{PE \leftarrow pre}$ , 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,

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

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,

$$\begin{aligned}\tau_V^{low} \cdot \dot{r}_V^{low} &= -r_V^{low} + (w_{V \leftarrow PE} \cdot \underline{r}_{PE}^{low})^2 = -r_V^{low} + (w_{V \leftarrow pPE} \cdot r_{pPE}^{low} + w_{V \leftarrow nPE} \cdot r_{nPE}^{low})^2 \\ \tau_V^{high} \cdot \dot{r}_V^{high} &= -r_V^{high} + (w_{V \leftarrow PE} \cdot \underline{r}_{PE}^{high})^2 = -r_V^{high} + (w_{V \leftarrow pPE} \cdot r_{pPE}^{high} + w_{V \leftarrow nPE} \cdot r_{nPE}^{high})^2.\end{aligned}\quad (4)$$

All values for neuron and network parameters, details on the model equations, 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,  $r_{\text{out}}$ , based on Bayesian multisensory integration (XXX),

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

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

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

## Inputs

The network receives feedforward stimuli  $s$  that may vary between trials. To account for noise, each stimulus is composed of  $N_{\text{in}}$  piece-wise constant values drawn from a normal distribution with mean  $\mu_{\text{in}}$  and standard deviation  $\sigma_{\text{in}}$ . To account for changes in the environment,  $\mu_{\text{in}}$  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 parameterisation 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 xxxx. Differential equations were numerically integrated using a 2<sup>nd</sup>-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

## Supporting Information

<b>A Detailed Methods</b>	<b>14</b>
A.1 Network model . . . . .	14
A.1.1 Prediction-error network model . . . . .	14
A.1.2 Memory and variance neuron . . . . .	15
A.1.3 Weighted output . . . . .	15
A.2 Connectivity . . . . .	16
A.2.1 Connections between neurons of the PE circuit . . . . .	16
A.2.2 Connections between the PE circuit and the M neuron . . . . .	16
A.2.3 Connections between the PE circuit and the V neuron . . . . .	17
A.3 Inputs . . . . .	17
A.4 Simulations . . . . .	18
<b>B Supporting analyses</b>	<b>19</b>
B.1 Activity of M and V neuron in a simplified model . . . . .	19
B.2 Impact of PE neurons' gain on estimating mean and variance . . . . .	20
B.3 Impact of PE neurons' baseline on estimating mean and variance . . . . .	20
B.4 Modelling the impact of neuromodulators on the sensory weight . . . . .	21
B.5 Sensory weight and contraction bias . . . . .	22
<b>C Supplementary Figures</b>	<b>22</b>

## 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 give 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 XXXref your own paperXXX.

#### 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 (?)

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

where  $\tau_E$  denotes the excitatory rate time constant ( $\tau_E=60$  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 ( $[•]_+$ ).

The dynamics of the activity of the dendritic compartments  $r_{nD}$  (nPE neuron) and  $r_{pD}$  (pPE neuron) obey (?)

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

where the weights  $w_{nD \leftarrow nE}$ ,  $w_{nD \leftarrow pE}$ ,  $w_{pD \leftarrow nE}$  and  $w_{pD \leftarrow pE}$  denote the recurrent excitatory connections between PCs, including backpropagating activity from the soma to the dendrites.  $w_{nD \leftarrow SOM}$  and  $w_{pD \leftarrow SOM}$  represent the strength of dendritic inhibition from the SOM neuron.  $w_{nD \leftarrow M}$  and  $w_{pD \leftarrow M}$  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,

$$\begin{aligned} r_X = [h_X]_+ \text{ with } \tau_I \frac{dh_X}{dt} &= -h_X + I_X + w_{X \leftarrow nE} \cdot r_{nE} + w_{X \leftarrow pE} \cdot r_{pE} + w_{X \leftarrow M} \cdot r_M - w_{X \leftarrow PV_1} \cdot r_{PV_1} \\ &\quad - w_{X \leftarrow PV_2} \cdot r_{PV_2} - w_{X \leftarrow SOM} \cdot r_{SOM} - w_{X \leftarrow VIP} \cdot r_{VIP}, \end{aligned} \quad (9)$$

where  $r_X$  denotes the firing rate of interneuron type  $X$ , and the weight  $w_{X \leftarrow Y}$  denotes the strength of connection between the presynaptic neuron  $Y$  and the postsynaptic neuron  $X$  ( $X \in \{PV_1, PV_2, SOM, VIP\}$ ,  $Y \in \{nPE, pPE, PV_1, PV_2, SOM, VIP, M\}$ ). The rate time constant  $\tau_I$  was chosen to resemble a fast GABA<sub>A</sub> 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,

$$\tau_E \cdot \frac{dr_M}{dt} = w_{M \leftarrow pE} \cdot r_{pE} - w_{M \leftarrow nE} \cdot r_{nE}. \quad (10)$$

$w_{M \leftarrow pE}$  denotes the connection strength between the pPE neuron and the memory neuron, and  $w_{M \leftarrow nE}$  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,

$$\tau_V \cdot \frac{dr_V}{dt} = -r_V + (w_{V \leftarrow pE} \cdot r_{pE} + w_{V \leftarrow nE} \cdot r_{nE})^2. \quad (11)$$

The weight  $w_{V \leftarrow pE}$  represents the connection strength between the pPE neuron and the variance neuron, while  $w_{V \leftarrow nE}$  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  $\tau_V$  was 5 s in the mean-field model, and 2 s in the population model (Fig. S2). XXX and 0.5 s in 2S3 – put everything for the not mean field model in the figure captions I would recommend, you can here say that it is only for the mean-field model ... XXX

### A.1.3 Weighted output

The weighted output  $r_{out}(t)$  is a linear combination of the current sensory input  $s(t)$  and the activity of the memory neuron,  $r_M(t)$ , inspired by Bayesian multisensory integration (XXX)

$$r_{out}(t) = \alpha \cdot s(t) + (1 - \alpha) \cdot r_M(t). \quad (12)$$

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

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

## 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 ???. 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 ( $g_{\text{nPE}}$  and  $g_{\text{pPE}}$ ) and the neuron numbers ( $N_{\text{nPE}}$  and  $N_{\text{pPE}}$ ):

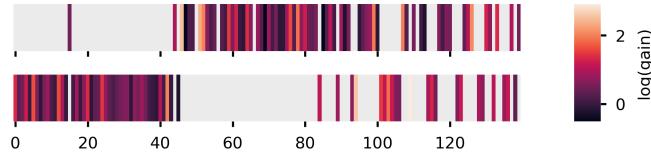
$$\begin{aligned}w_{M \leftarrow nE} &= \frac{-\lambda}{g_{\text{nPE}} \cdot N_{\text{nPE}}} \\ w_{M \leftarrow pE} &= \frac{\lambda}{g_{\text{pPE}} \cdot N_{\text{pPE}}}\end{aligned}$$

where  $\lambda$  denotes the speed at which the perfect integrator evolves (lower PE circuit:  $\lambda = 3 \cdot 10^{-3}$  for the mean-field model in Fig. 2,  $\lambda = 5 \cdot 10^{-2}$  for the population model in Fig. S2, and  $\lambda = 4.5 \cdot 10^{-2}$  otherwise; higher PE circuit:  $\lambda = 7 \cdot 10^{-4}$ ). XXX in Fig. S3 wPE2PE \* 100 =; I think it is best to put this is the captions, I mean for 2S2 and 2S3 ... for mean field network can be shown here and said "if not stated otherwise in the figures/captions"

For the mean-field networks ( $N_{\text{nPE}} = N_{\text{pPE}} = 1$ ), 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 $\rightarrow$ SOM FB $\rightarrow$ VIP	MFN 2 FB $\rightarrow$ SOM FF $\rightarrow$ VIP	MFN 3 FF $\rightarrow$ SOM FF $\rightarrow$ 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 ???. 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  $w_{X \leftarrow M}$ . If a connection exists, we assume  $w_{X \leftarrow M} = 1$ . 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.

<b>Network</b>	<b>MFN 1</b>	<b>MFN 2</b>	<b>MFN 3</b>
	FF → SOM FB → VIP	FB → SOM FF → VIP	FF → SOM FF → VIP
SOM	0	1	0
VIP	1	0	0

**Table 2.**  $w_{X \leftarrow M}$  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,

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

We introduce the factor  $\theta$  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  $nPE \cdot pPE$  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 XXX, each mean-field network receives a stimulus  $s_i$  drawn from a distribution. In this case,  $\theta$  must be chosen such that deviations from the true variance can be mitigated or fully corrected. The true  $\theta$  depends on the distribution at hand. In our simulations, we used a uniform distribution  $U(a, b)$ , in which case  $\theta$  can be derived from

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

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  $r_{pE} = r_{pD} = r_{nE} = r_{nD} = 0 \text{ s}^{-1}$  and  $r_P = r_S = r_V = 4 \text{ s}^{-1}$ . In the case of the population network, we set the external inputs of all neuron types to  $5 \text{ s}^{-1}$ , 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  $w_i = 1 - w_{X \leftarrow M}$  (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 XXXX. Differential equations were numerically integrated using a 2<sup>nd</sup>-order Runge-Kutta method. Neurons were initialized with  $r = 0/\text{s}$ .

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 ???. 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. ???. 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 trial, 10 constant values were drawn from a normal distribution  $N(\mu_{in}, \sigma_{in}^2)$ , each 500 time steps long. The stimulus mean was drawn from an uniform distribution,  $U(a, b)$ , with mean  $\mu_{in}$  and variance  $\sigma_{trial}^2$ .

**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  $\gamma$  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:  $\sigma_{in}^2 = 0$  and  $U(1, 9)$ , D:  $\sigma_{in}^2 = 5$  and  $U(5, 5)$ , F:  $\sigma_{in}^2 = 0$  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 trials. State 1: stimulus  $\in N(\mu_{in}, 0)$  with  $\mu_{in} \in U(5, 5)$ , State 2: stimulus  $\in N(\mu_{in}, 3)$  with  $\mu_{in} \in U(5, 5)$ , State 3: stimulus  $\in N(\mu_{in}, 3)$  with  $\mu_{in} \in U(0, 10)$ , State 4: stimulus  $\in N(\mu_{in}, 0)$  with  $\mu_{in} \in U(0, 10)$ .

**Figure 3 Supplementary Fig. 2:**  $\mu_{in} = 5$  and  $\sigma_{trial}^2 \in \{0, 0.75, 1.5, 2.25, 3\}$ ,  $\sigma_{in}^2 \in \{3, 2.25, 1.5, 0.75, 0\}$ . A: scaling factors of  $w_{M \leftarrow PE}$  were 0.3 and 7.

**Figure 4:**  $\mu_{in} = 5$ . A, top:  $\sigma_{in}^2 = 0$  and  $\sigma_{trial}^2 = 1$ . A, bottom:  $\sigma_{in}^2 = 1$  and  $\sigma_{trial}^2 = 0$ . C:  $\sigma_{in}^2 = 1$  and  $\sigma_{trial}^2 = 1$ . D:  $\sigma_{in}^2 = 5$  and  $\sigma_{trial}^2 = 0$ . 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 (see Eqs. XXX). The results (left) were computed for baselines  $\in [0, 3]$ , while the results (right) were computed for gains  $\in [0.5, 1.5]$ . XXX Here refer to equations XXX.

**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  $\sigma_{\text{in}}$  is a linear function of  $\mu_{\text{in}}$ . 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, and compare our network with the Kalman filter as well as Bayes factor surprise. XXX

### 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

$$\begin{aligned} r_{\text{nE}} &= [r_M - s_{\text{FF}}]_+ \\ r_{\text{pE}} &= [s_{\text{FF}} - r_M]_+ \end{aligned} \quad (15)$$

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

$$\tau_M \cdot \frac{dr_M}{dt} = r_{\text{pE}} - r_{\text{nE}} \quad (16)$$

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

$$\tau_M \cdot \frac{dr_M}{dt} = -r_{\text{nE}} = -r_M + s_{\text{FF}}. \quad (17)$$

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

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

Hence, the activity of  $r_M$  is given by

$$r_M = \frac{1}{\tau_M} \int_0^t e^{-(t-x)/\tau_M} \cdot s_{\text{FF}}(x) \, dx \quad (18)$$

for zero activity at time  $t = 0$ . In the limit of  $t \rightarrow \infty$  (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

$$\tau_V \cdot \frac{dr_V}{dt} = -r_V + (r_{\text{pE}} + r_{\text{nE}})^2 = -r_V + (r_M - s_{\text{FF}})^2, \quad (19)$$

leading to the time-dependent solution

$$r_V = \frac{1}{\tau_V} \int_0^t e^{-(t-x)/\tau_V} \cdot [r_M(x) - s_{\text{FF}}(x)]^2 \, dx. \quad (20)$$

In the limit of  $t \rightarrow \infty$ ,  $r_V$  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.

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

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

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

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

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

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

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

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

Inserting eqs. (22) yields

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

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

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

$n_0$  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

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

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

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

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

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

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

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

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 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,

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

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 a number of 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 is 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

The contraction bias can be formalized through the slope in a linear equation describing the relationship between the estimated and true sensory stimulus. If the slope is close to 1, the bias is small. In contrast, if the slope is close to zero, the bias is large.

If the prediction encoded in the activity of the M neuron changes very slowly, the bias is well-captured by the sensory weight,

$$r_{\text{out}} = \alpha_S \cdot s_{\text{FF}} + \text{constant} \quad (31)$$

However, the prediction is normally a function of the history of sensory stimuli the network received, and cannot be approximated by a constant. To investigate how the slope (describing the contraction bias) depends on the sensory weight and other factors, let us resume a toy model in which we assume that the prediction decays exponentially with time constant  $\tau$  to a presented constant stimulus value,  $s$ ,

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

with  $P_0$  describing the initial value at time  $t = 0$ . Let us 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  $n$ th stimulus value can be expressed as

$$P_n = P_0 \cdot e^{-\Delta t/\tau_M} + (1 - e^{-\Delta t/\tau_M}) \sum_{i=1}^n s_i \cdot e^{-(n-i) \cdot \Delta t/\tau_M} \quad (33)$$

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

$$\langle P_n \rangle_s = e^{-\Delta t/\tau_M} \int_{\mu - \frac{\sigma_P}{12}}^{\mu + \frac{\sigma_P}{12}} P_0 f(P_0) dP_0 + (1 - e^{-\Delta t/\tau_M}) \sum_{i=1}^n e^{-(n-i) \cdot \Delta t/\tau_M} \int_{s - \frac{\sigma_S}{12}}^{s + \frac{\sigma_S}{12}} x f(x) dx \quad (34)$$

Solving the integrals yield

$$\langle P_n \rangle_s = \mu \cdot e^{-T/\tau_M} + (1 - e^{-\Delta t/\tau_M}) \sum_{i=1}^n e^{-(n-i) \cdot \Delta t/\tau_M} \cdot s \quad (35)$$

Making use of the geometric series, the expression simplifies to

$$\langle P_n \rangle_s = \mu \cdot e^{-T/\tau_M} + (1 - e^{-T/\tau_M}) \cdot s$$

Inserting the expression in the equation for the weighted output yields

$$r_{\text{out}} = [\alpha_S e^{-T/\tau_M} + (1 - e^{-T/\tau_M})] \cdot s + (1 - \alpha_S) e^{-T/\tau_M} \mu$$

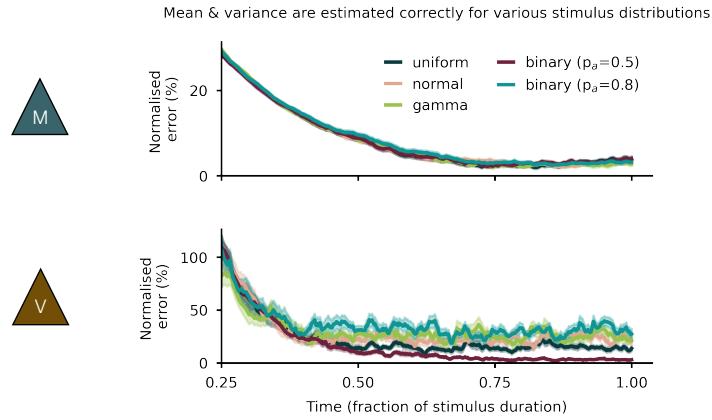
Hence, the slope  $m = [\alpha_S e^{-T/\tau_M} + (1 - e^{-T/\tau_M})]$  depends on the sensory weight  $\alpha_S$ , the trial duration  $T$  and time constant  $\tau$ . Please note that the sensory weight is a function of the trial duration itself (see Fig. 3F). However, for illustration purposes, we take  $\alpha_S$  to be constant here.

In this toy model, if the variance of the prediction is zero (that is, in a prediction-driven input regime),  $\alpha_S \approx 0$ , and, hence, the slope is  $(1 - e^{-T/\tau_M})$ . Thus, decreasing or increasing the stimulus variance does not have an effect on the bias (see Fig. 5 D).

Likewise, if the variance of the sensory stimulus is zero (that is, in a stimulus-driven input regime),  $\alpha_S \approx 1$ , and, hence, the slope is 1. Thus, decreasing or increasing the trail variance does not have an effect on the bias (see Fig. 5 C).

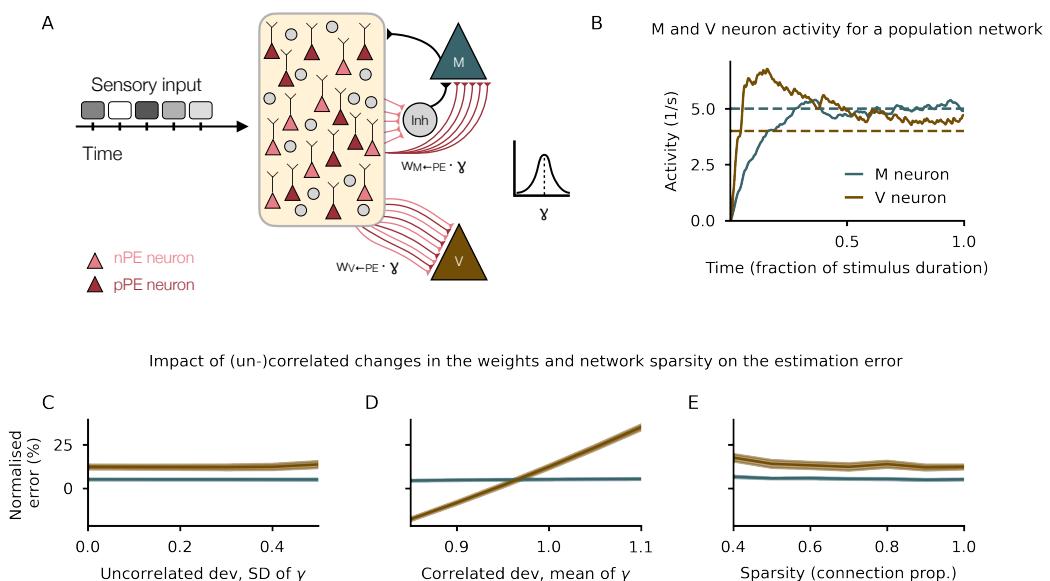
Moreover, the toy model also reveals that if  $T \rightarrow \infty$ , the slope  $m$  goes to 1, suggesting that without further noise in the system, the network can perfectly represent the sensory stimulus (the prediction would approach the current stimulus).

## C Supplementary Figures



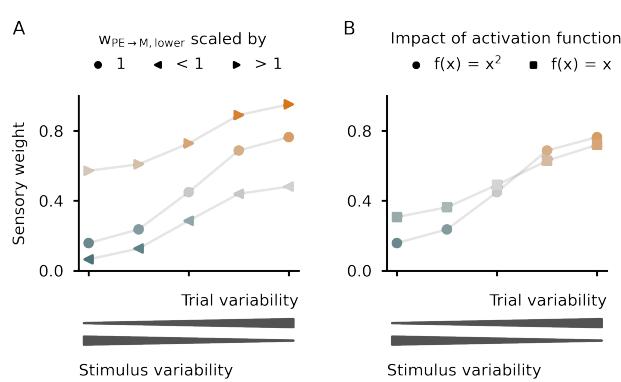
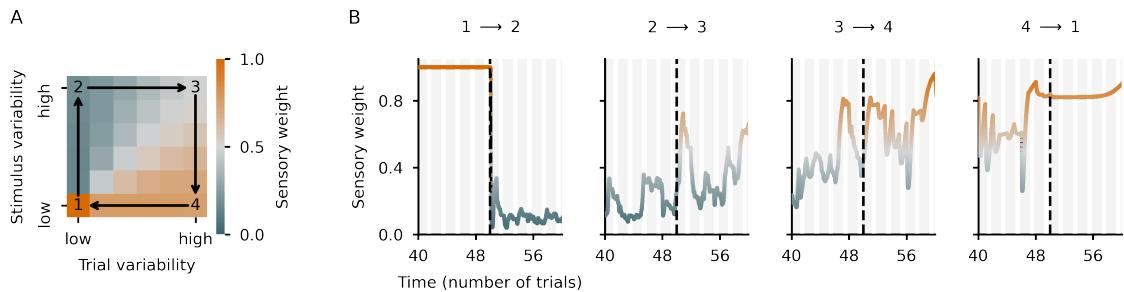
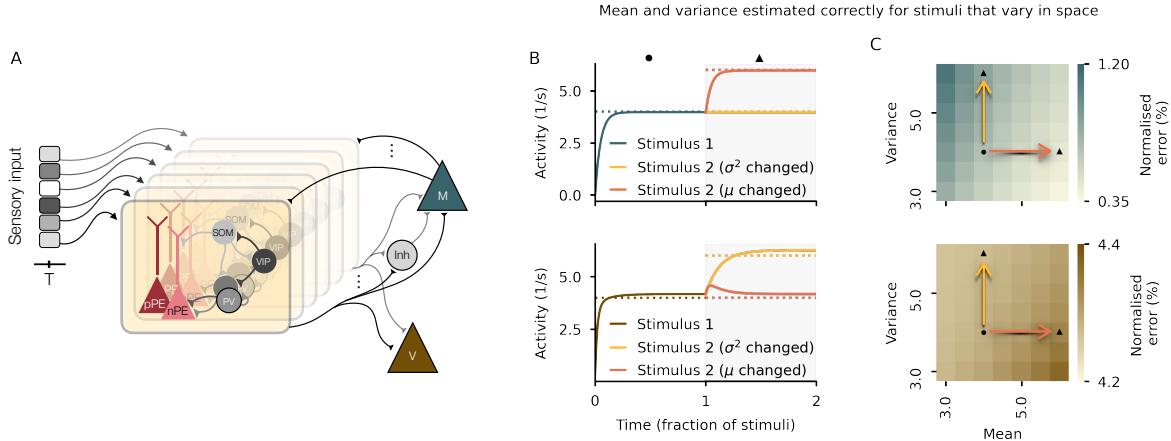
**Figure S1.** Estimating mean and variance of different stimulus distributions.

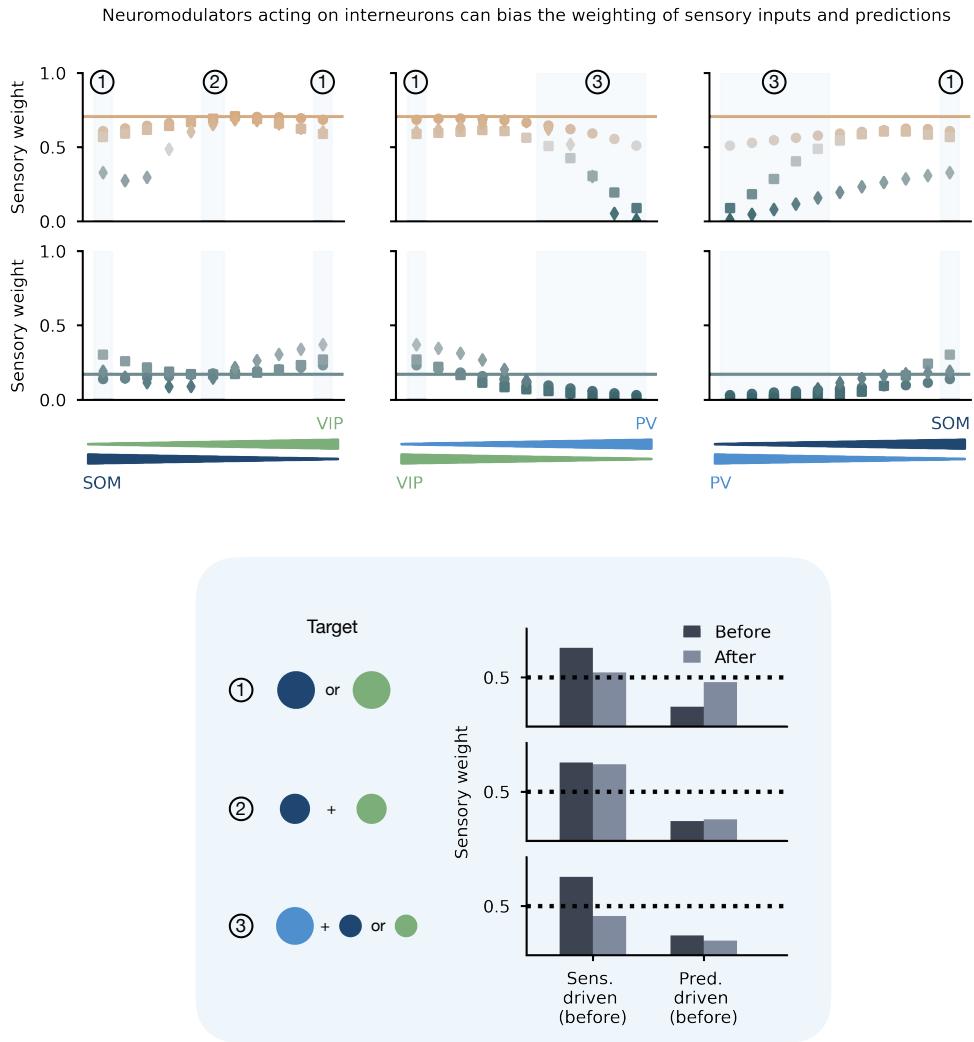
Top: The normalised absolute difference between the averaged mean and the activity of the M neuron decreases to a near-zero level for all stimulus distributions tested. Bottom: The normalised absolute difference between the averaged variance and the activity of the V neuron decreases with small differences 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.

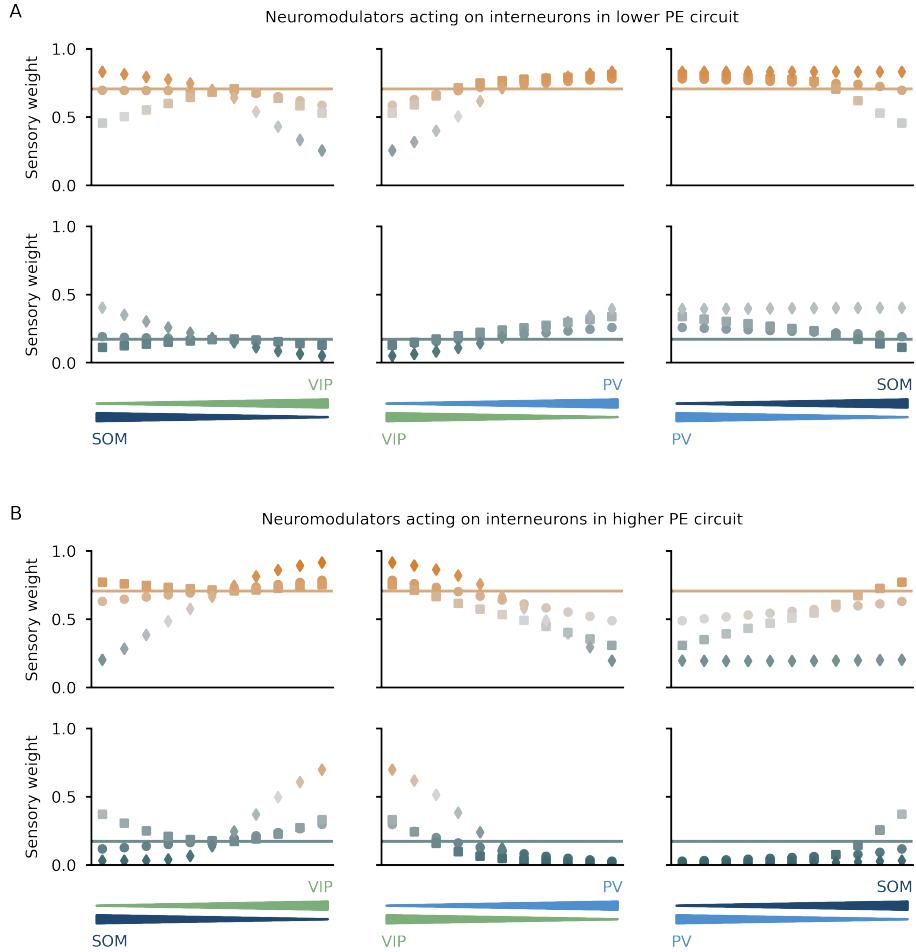
(A) Illustration of the rate-based population network and the stimuli over time. (B) M and V neuron activities over time for one example parameterisation. (C) The normalised absolute difference between the averaged mean and the activity of the M neuron (dark green) or between the averaged variance and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





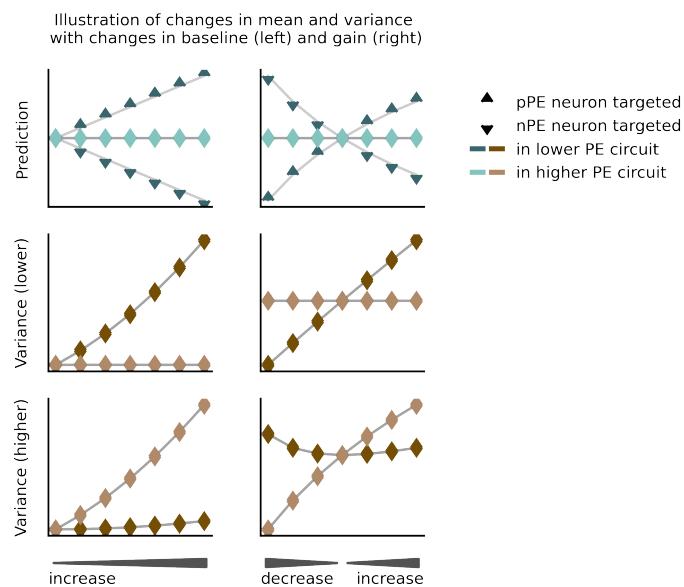
**Figure S6. The impact of neuromodulators acting globally on groups of interneurons.**

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).



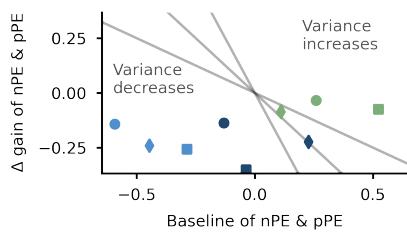
**Figure S7. The impact of neuromodulators acting locally on groups of interneurons.**

(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 XXX, the contribution of gain and baseline to the changes in the mean and variance estimation can be revealed. XXX Equations



**Figure S9.** The combined changes in baseline and gain of all PE neurons determine the shift in the sensory 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.