Hierarchy of prediction errors shapes the learning of context-dependent sensory representations

Matthias C. Tsai[1], Jasper Teutsch[2,3,4], Willem A.M. Wybo[5], Fritjof Helmchen[6,7,8], Abhishek Banerjee[2,3*†], Walter Senn[1*†]

[1]Department of Physiology, University of Bern, Bern, Switzerland.
[2]Department of Pharmacology, University of Oxford, Oxford, UK.
[3]Blizard Institute, Queen Mary University of London, London, UK.
[4]Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK.
[5]Peter Grünberg Institute, Jülich Research Center, Jülich, Germany.
[6]Brain Research Institute, University of Zürich, Zürich, Switzerland.
[7]Neuroscience Center Zürich, University of Zürich, Zürich, Switzerland.
[8]University Research Priority Program (URPP), Adaptive Brain Circuits in Development and Learning (AdaBD), University of Zürich, Zürich, Switzerland.

*Corresponding author(s). E-mail(s): abhishek.banerjee@pharm.ox.ac.uk; walter.senn@unibe.ch;
Contributing authors: matthias.chinyen.tsai@gmail.com; jasper.teutsch@pharm.ox.ac.uk; w.wybo@fz-juelich.de; helmchen@hifo.uzh.ch;
†These authors contributed equally to this work

Abstract

How sensory information is interpreted depends on context. Yet, how context shapes sensory processing in the brain, remains elusive. To investigate this question we combined computational modeling and *in vivo* functional imaging of cortical neurons in mice during reversal learning of a tactile sensory discrimination task. During learning, layer 2/3 somatosensory neurons enhanced their response to reward-predictive stimuli, explainable as gain amplification from apical dendrites. Reward-prediction errors were reduced and confidence in the outcome prediction increased. Upon rule-reversal, the lateral orbitofrontal cortex, through disinhibitory VIP interneurons, encoded a context-prediction error signaling a loss of confidence. The hierarchy of prediction errors in cortical areas is mirrored in top-down signals modulating apical activity in the primary sensory cortex. Our model explains how contextual changes are detected in the brain and how errors in different cortical regions interact to reshape and update the sensory representation.

Keywords: top-down modulation, representation learning, lateral orbitofrontal cortex, reversal learning, apical amplification, primary somatosensory cortex, continual learning, non-stationary environment, interneurons, disinhibition, online machine learning

The theory of reward-prediction errors with their neuronal correlates represents a fundamental milestone for behavioural neuroscience [1]. It forms the foundation for reinforcement learning in the brain and inspired developments in artificial intelligence [2, 3]. Making reward predictions in complex environments, however, also requires the correct identification of ongoing contexts. For one context, the rewards in an environment might follow certain rules; in another context these rules might be altered, hence altering the reward contingency. Yet, it is not clear – from a normative point of view – how context is represented, how a change in context is even detected, and how a putative context change can reshape computation. Although the ultimate goal of reward-based learning lies in minimizing reward-prediction errors and selecting actions to maximize reward, computing context-prediction errors and reducing them constitutes an essential intermediary goal. A complete model of the brain should therefore explain how reward- and context-prediction errors jointly modulate decision making from the early steps of sensory processing.

Being able to extract a broad range of features from the sensory world is crucial to make accurate decisions in varying contexts. However, having a rich sensory representation also enlarges the search space from which to identify

behaviorally relevant information. The brain addresses this dilemma by selectively amplifying task-relevant sensory responses [4–6] (Fig. 1a). While the origin of these enhanced responses in the sensory cortex is not fully understood, they can be modeled as multiplicative gain amplification [7]. Sensory response amplification depends on context, which indicates a top-down attentional modulation rather than a bottom-up adaptation [8–10]. Apical dendrites of pyramidal neurons receive top-down afferents [11], multiplicatively modulate the firing rate [12, 13] and acquire strengthened responses to stimuli that predict upcoming salient outcomes [14, 15]. Therefore, apical dendrites are suspected to selectively amplify neural responses in order to increase the internal salience of task-relevant sensory stimuli [16, 17].

The amplification of a select subset of features, however, can have further drawbacks. A once advantageous salience allocation can become inadequate after a change in context. Therefore, detecting such contextual changes becomes crucial for the adaptability of behavior. In animals, such scenarios are studied by means of reversal learning tasks [18]. Most prominently, lateral orbitofrontal cortex (lOFC) has been shown to respond to changes in reward contingency and plays a critical role in flexibly adjusting behavior [18, 19]. Further work has shown that the lOFC can inhibit apical dendrites of pyramidal neurons in the visual cortex (V1) via somatostatin-expressing (SST) interneurons [20], which could be a potential mechanism to revert acquired apical amplification patterns that are no longer appropriate after a change in context.

Here, we measured functional responses from S1 and lOFC neurons of mice during reversal learning and integrated our findings into a computational model explaining how the lOFC can support behavioral flexibility by actively reshaping sensory representations. We examined the robustness of online reinforcement learning in the presence of a large set of irrelevant stimuli and show that the reward-guided apical gain amplification of sensory pyramidal neurons enhanced performance. We show that, upon rule reversal, lOFC signals a context-prediction error during the stimulus presentation as a consequence of observing highly unexpected outcomes. The context-prediction error inhibits the sensory top-down amplification learned during the previous rule, and allows for the behavioral adaptation to the new rule. Finally, we also demonstrate that silencing vasoactive intestinal polypeptide-expressing (VIP) interneurons in the lOFC impairs rule-switching. Taken together, our results show how the brain builds context-dependent sensory representations to support behavioral adaptation in non-stationary environments, and provide new perspectives on the roles of distinct inhibitory mechanisms in that process.

Results

Value-based apical amplification of sensory representations in S1

Learning from large sensory representations containing few task-relevant and many task-irrelevant neurons poses a challenge. The task-irrelevant neurons will act as distractors and obfuscate which signals to base decisions on. This applies at a high level, e.g., animals perceiving with multiple senses to navigate complex environments. However this issue grows considerably as the brain extracts diverse features from each sensory stimulus (Fig. 1a). We replicated this problem for a rewarded go/no-go sensory discrimination task, where action selection was learned with online policy gradient. The occurrence of a stimulus s_1 could lead to a reward if a specific action (go) was performed, and another stimulus s_2 led to a punishment that could be avoided by not performing that go-action (Fig. 1b). The sensory representation was made of rate-based neurons with two neurons each responding to a stimulus (s_1 or s_2) and varying numbers of randomly activated distractor neurons (Fig. 1c). In addition, the neuronal activity was modulated by a multiplicative gain. Different simulations varied regarding the gain applied on the s_1 and s_2 selective neurons, while the gains of the distractor neurons remained fixed to 1 (Fig. 1d). For equal gains, the task was learned consistently in the presence of few distractors, but the learning performance began to falter around 100 distractor neurons (Fig. 1e). Differences in gain of the task-relevant neurons compared to the distractors also affected learning, i.e., weak task-relevant signals could spoil the performance. However, the performance degradation caused by an overwhelming number of distractors could be rescued by increasing the gain of the s_1 and s_2 stimuli encoding neurons (Fig. 1f).

In summary, sensory representations can be improved by amplifying the response of task-relevant sensory neurons, which experimentally also correlates with behavioral performance [5, 21]. At the same time, apical activity has been shown to increase the somatic excitability [22]; learning enhances the apical response to behaviorally relevant stimuli [14, 15]; and adaptation in apical dendrites of sensory neurons precede behavioral learning [23]. We, therefore, postulated that changes in sensory responses promoting task-relevant information could be modeled as top-down excitation of apical dendrites that multiplicatively amplify bottom-up basal activations (Fig. 1d and Extended Data Fig. S1a). The guiding principles underlying this conditioning of the sensory representation remain unclear, but the enhancement of responses to stimuli that are predictive of future rewards has been reported with high consistency [18, 21, 24–26] (Extended Data Fig. S2a).

We replicate the experimental observations by modeling pyramidal neurons that modulate bottom-up sensory inputs with top-down signals encoding the stimulus-reward association (Fig. 1d). A reward prediction network derives neuron-specific values that measure whether bottom-up inputs are associated with future rewards (Fig. 1g). The sensory neurons are then amplified by shaping apical excitation according to this neuron-specific value, which encodes

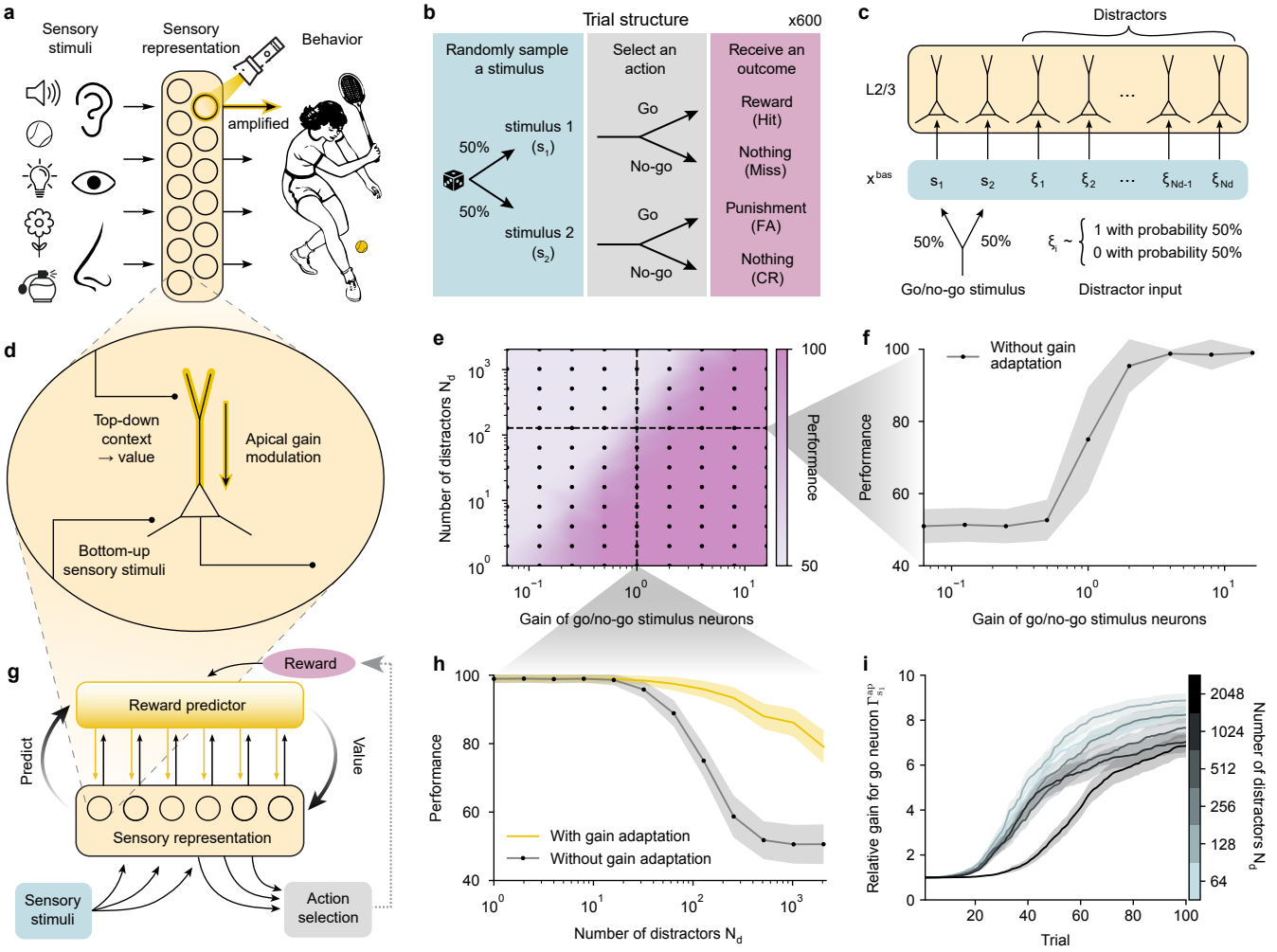

Fig. 1 Value-based apical gain modulation amplifies task-relevant sensory signals. **a**, Sensory representations contain information about many sensory inputs and selectively amplify features that are most important for behavioral decisions. **b**, Structure of the go/no-go sensory discrimination task with each combination of stimulus and action leading to an outcome (Hit, receive reward; CR, correct rejection, omission of punishment; FA, false alarm, receive punishment; Miss, omission of reward). **c**, Bottom-up inputs to the sensory pyramidal neurons. The go (s_1) and no-go (s_2) stimuli drive separate neurons. Task-irrelevant features drive distractor neurons. **d**, Schematic showing how top-down afferents onto pyramidal neurons multiplicatively modulate the gain of the bottom-up input. In the model, the apical input reflects the neuron-specific value and the bottom-up input has a fixed selectivity. **e**, Performance of the online policy gradient algorithm without gain adaptation on a go/no-go sensory discrimination task depending on the number of distractor neurons and on the gain of the two sensory neurons selective for either s_1 or s_2. The gain of the distractor neurons was always set to 1. Performance was evaluated after 600 trials of training. Each black dot corresponds to the average from 64 different pseudorandom initializations. **f**, Horizontal slice of panel e with 128 distractor neurons and different fixed gains for the go and no-go stimulus encoding neurons. The grey area shows the standard deviation. **g**, Computational model of sensory processing. Bottom-up sensory stimuli drive the pyramidal neurons of the sensory representation. The bottom-up activation is amplified by top-down apical afferents conveying neuron-specific value. The values, derived from a reward prediction network, measure how much the sensory neuron activities are associated with future rewards. The sensory pyramidal neurons project both to the reward prediction network and to a policy network that selects actions. **h**, Performance achieved with and without adaptive gain modulation depending on the number of distractor neurons. The values without gain adaptation correspond to the vertical slice in panel e. **i**, As learning progressed, the apical gain of the go-stimulus (s_1) selective neuron increased, relative to the average gain of the other neurons, see equation (12). The mean and the standard error of this relative gain Γ^{ap}_{s1} are shown for simulations with different numbers of distractors across the first 100 trials. With greater numbers of distractor neurons, the difficulty to identify reward-associated neurons increases, thus Γ^{ap}_{s1} requires more trials to rise.

the expected relevance of each neuron for predicting future rewards (see Supplementary Equations 2). The simulated learning performances confirm that reward-guided apical gain modulation renders the system more robust to the presence of task-irrelevant stimuli(Fig. 1h). The resulting apical salience map increases the responses of sensory neurons according to their reward-association (Fig. 1i), and this in turn improves the sensory representation for action selection. Hence, top-down projections onto apical dendrites of sensory neurons support behavioral learning by amplifying task-relevant information and increasing the signal-to-noise ratio (Extended Data Fig. S3).

Lateral OFC signals a context-prediction error at stimulus time

So far, the model made use of reward predictions in a fixed stationary context. Next, we show how tracking the second order statistics of reward-prediction errors is helpful in non-stationary environments, e.g., in case the rule of the go/no-go task is reversed (Fig. 2a). When being exposed to a new task, animals may be agnostic about their expectations, and any outcome can be considered equally uncertain. However, as an internal model of the task is constructed and the expected outcomes become more and more accurate, the uncertainty in the reward prediction decreases. At the same time, any deviation from the predicted outcome becomes increasingly unexpected. How expected or unexpected an outcome is, can be captured by using an estimate $\hat{\sigma}^2$ of the squared reward-prediction error σ^2. The estimate $\hat{\sigma}^2$ evaluates the variance unexplained by the model (see Supplementary Equations 4) and can be interpreted as the expected uncertainty about reward predictions (see Supplementary Equations 3). As learning progresses and the reward predictions $\hat{Q}$ improve, $\hat{\sigma}^2$ will decrease (Fig. 2b). In consequence, a reward after the go-stimulus (s_1) will be predicted with increasing confidence, while any different outcome becomes less and less expected.

The confidence c in the reward prediction is defined as a monotonically decreasing function of the prediction uncertainty $\hat{\sigma}^2$, see equation (7). It is an estimate of the squared fraction of variance explained (FVE^2). The higher the FVE, i.e. the more variance in the observed outcomes can be predicted by the model, the higher is the confidence c in that model and its outcome predictions. After learning, if a rule reversal takes place (Fig. 2a), highly unexpected outcomes will be observed (Fig. 2b). This leads to an increase in the uncertainty ($\Delta\hat{\sigma}^2 > 0$), and hence a decrease in the confidence about the prediction ($\Delta c < 0$, Fig. 2c). The confidence loss (CL) provides an internal signal, suggesting to the animal that a change of context has occurred. The CL is elicited when unexpected outcomes are observed.

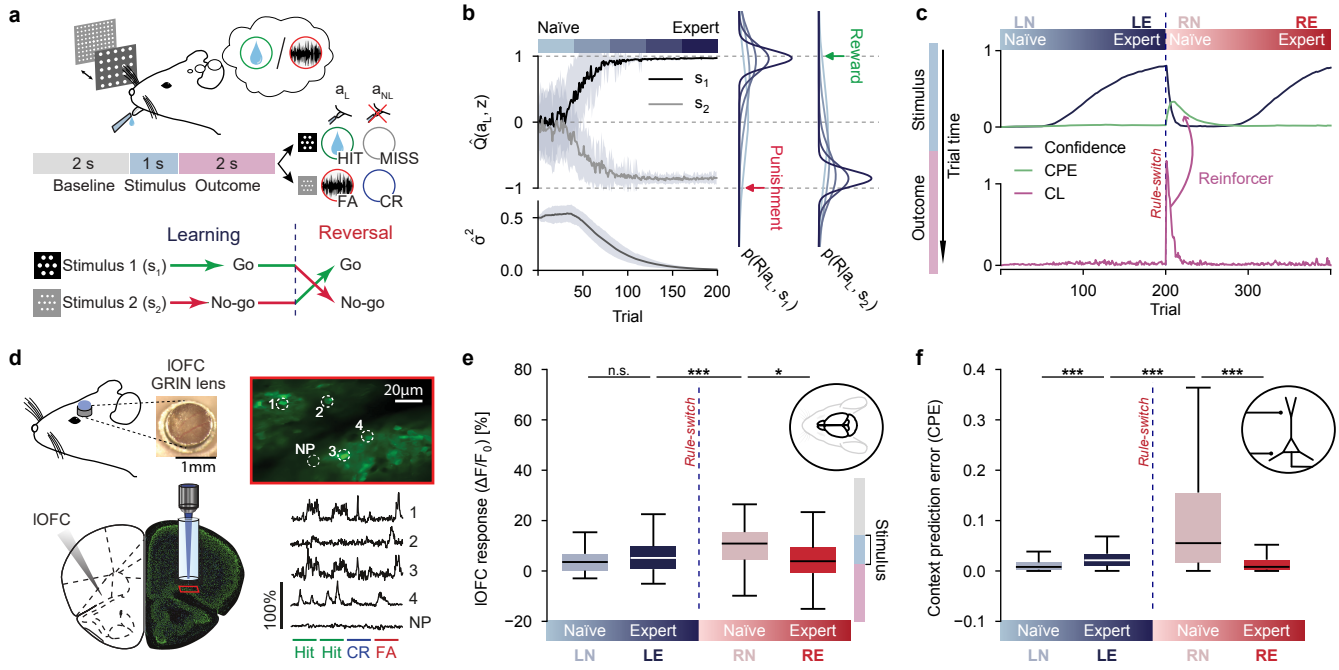

Fig. 2 Lateral orbitofrontal cortex signals a context-prediction error at stimulus time. a, Go/no-go sensory discrimination task with rule reversal. Top: schematic of experimental setup, trial structure and possible outcome types (Hit, CR, FA, Miss). Bottom: after animals show stable expert discriminatory performance (learning), stimulus-outcome contingencies are reversed (reversal). **b**, Evolution of the reward prediction $\hat{Q}$ depending on the sensory representation z and the lick-action a_L. Top: the values $\hat{Q}(a_L, z)$ are shown across trials for the rewarded go-stimulus (s_1) and for the punished no-go-stimulus (s_2). Bottom: The estimate of the squared reward-prediction error ($\hat{\sigma}^2$) tracks the uncertainty of the reward prediction. Right: Estimated probability densities for the predicted reward as Gaussian distributions parameterized by the mean $\hat{Q}(a_L, z)$ and variance $\hat{\sigma}^2$ for either stimulus (s_1 or s_2). As learning progresses from the naive (light blue) to the expert stage (dark blue), the estimated densities become more accurate and narrow. Colored arrows indicate unexpected outcomes in case of rule reversal, such as punishment following s_1 (red) or reward following s_2 (green). **c**, Top: Prediction confidence (black) and context-prediction error (CPE, green) during the stimulus window, see equations 7-9). Bottom: The confidence loss (CL) triggered by unexpected outcomes. The CL entrains the CPE to appear before the outcome is observed (purple 'Reinforcer' arrow pointing backwards in trial time). **d**, Left: schematic and photograph of LOFC cannula and GRIN lens implantation. Right, two-photon fluorescence image of L2/3 lOFC neurons imaged through the GRIN lens and extracted traces of example cells and neuropil (NP). Traces of four concatenated example trials (whole trial) and corresponding trial outcome from an expert animal (LE). **e**, Box plots of Ca^{2+} transient amplitudes ($\Delta F/F_0$) from L2/3 lOFC neurons in the stimulus window during different learning phases of the reversal learning task (39, 122, 166, 170 neurons from 3 mice during the LN, LE, RN, RE phase respectively). The phases correspond to learning naive (LN), learning expert (LE), reversal naive (RN) and reversal expert (RE). Statistical differences were tested with two-tailed, independent samples T-tests with Bonferroni corrections, the asterisks indicate whether the p-values were > 0.05 (n.s), ≤ 0.05 (*), ≤ 0.01 (**) or ≤ 0.001 (***). This convention is maintained throughout the manuscript. **f**, Same as in panel e, for the simulated context-prediction error. Note: panels e and f use pictograms of a mouse head and neuron model to distinguish experimental and simulation data. This convention is maintained throughout the manuscript.

Yet, to affect behavior, the knowledge that the context has changed should take effect earlier within trial time, when sensory signals are processed and actions are decided. Neurons in the lOFC have been shown to respond to unexpected outcomes following a rule reversal [18, 27, 28]. We therefore explored whether lOFC also produced temporally advanced anticipatory signals informing that a rule reversal has occurred.

Analogously to the reward-predictor used to model top-down afferents of sensory neurons, we modeled the lOFC activity to become predictive for the confidence loss triggered by unexpected outcomes. In consequence, the anticipatory signal in the stimulus window predicts the confidence loss in the outcome window. Since it is reinforced by steady CL from successive unexpected outcomes, we refer to this temporal predictor of the CL as *context-prediction error* (CPE). Because the CPE is predictive for the CL that is itself based on the second-order statistics ($\hat{\sigma}^2$) of the reward-prediction error, the CPE can be viewed as a second-order reward-prediction error. Accordingly, within few trials following a rule reversal, the CPE produced a robust internal signal of a contextual change — in this case of the fact that the rule was switched (Fig. 2c). To test our model we then trained mice in a go/no-go texture discrimination task containing a rule reversal (Fig. 2a), and recorded the activity of L2/3 lOFC neurons with calcium imaging (Fig. 2d). We found that the lOFC response during the stimulus window increases in the trials closely following reversal, which is in accordance with a neural population encoding the CPE (Fig. 2f). We will therefore study how the lOFC might influence learning and behavior by modeling the lOFC population with the CPE.

Lateral OFC cancels top-down amplification in S1 after rule reversal

In scenarios where the context can change, the amplification of sensory features conditioned during a given context can be ill-suited in another context. Since lOFC neurons can signal changes in context prospectively at stimulus time (Fig. 2e), it raises the question of whether the lOFC could help to reshape top-down amplification at the apical dendrites. SST interneurons are prominent inhibitors of apical dendrites [29, 30]. Experimental findings have also shown that post-learning reactivation of SST could restore naïve-like activity patterns [31] and that lOFC can inhibit apical dendrites of pyramidal neurons via SST interneurons in V1 [20]. We therefore presume that lOFC projects to SST interneurons in S1 to counteract context-dependent top-down excitation (Fig. 3a and Extended Data Fig. S5). Doing so, strong lOFC responses during the stimulus windows following a rule reversal will inhibit previously conditioned sensory amplification patterns and allow the new reward-associated stimulus to be amplified.

Our Ca^{2+}-imaging data of L2/3 excitatory S1 neurons shows that the enhanced response to the rewarded stimulus s_1, acquired during learning of the first rule, moves to the newly rewarded stimulus s_2 after learning the reversed rule (Fig. 3b). When chemogenically silencing lOFC neurons after the rule reversal (via CNO-activated hM4Di, see Methods), no significantly enhanced response to the newly rewarded stimulus s_2 was developed (Fig. 3b). The evolution of the S1 response and its dependence on lOFC could be reproduced in the model by including a lOFC feedback to SST interneurons of the sensory cortex (Fig. 3c). With this inhibitory feedback, the selective top-down amplification of reward-predictive sensory stimuli can be reverted by lOFC, provided that a context change leads to a strong context-prediction error (Fig. 3d).

Besides measuring the S1 responses, we also compared the performances achieved by the model with and without lOFC after reversal. We found that the model without lOFC failed to adapt to the reversed rule (Fig. 3e). However, if the context-prediction error from lOFC drives apical inhibition in S1, the incorrect amplification of sensory neurons from bygone contexts was prevented. In consequence, by unbiasing the sensory representation, lOFC allows the reversed rule to be learned almost as fast as the original rule (Fig. 3f). Interestingly, animals with silenced lOFC can eventually adapt to the reversed rule after significantly more trials [18], which the model also reproduced (Fig. 3g).

Context-prediction errors support learning for multiple reversals

Besides unbiasing the sensory representation, inhibiting the top-down sensory amplification upon rule reversal also directly affects action selection (Fig. 4a). The conditioned behavior is inhibited at the source, by reducing the response to the stimuli responsible for triggering the behaviour in the first place. At the same time, reducing sensory neuron responses also decreases the plasticity of their efferent synapses (see Supplementary Equations 1). As a consequence, the synapses encoding the policy for the original rule are protected from degradation. Therefore, behavior can be adapted to the reversed rule, while still preserving the synaptic weights encoding the sensorimotor mapping for the original rule (Fig. 4c and Extended Data Fig. S7). Upon a second rule reversal, we observe that the behavior adjusts faster than during the first reversal (Fig. 4d). This can be attributed to the protection of the sensorimotor mapping acquired in the first learning stage. Indeed, behavioral adaptation requires both an appropriate sensorimotor mapping and sensory top-down modulation [6] (Fig. 4a). As only the latter needs to be adjusted after the second reversal (Fig. 4e), the overall behavioral adaptation can occur more rapidly.

By preserving the sensorimotor mappings of prior rules, the appropriate behavior for these rules can be restored, merely by recovering the top-down sensory amplification pattern. Until now we considered top-down excitation of sensory neurons to be driven by a single, fixed context representation vector. Given the single context representation, the top-down projections had to relearn the contextual value of each neuron upon each rule switch (Fig. 4c). To that

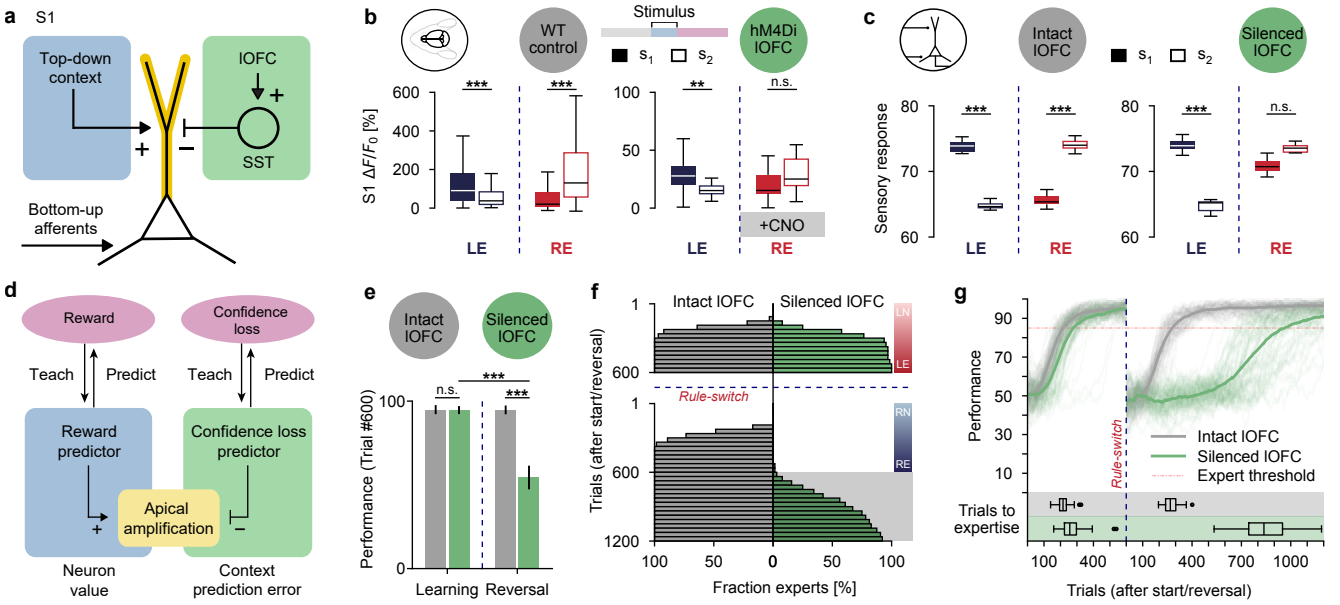

Fig. 3 Lateral OFC inhibits the top-down amplifications in S1 upon reversal. a, Apical dendrites of the S1-pyramidal neurons (yellow) are excited by top-down contextual afferents (blue) and inhibited by the lOFC (encoding CPE) via SST interneurons (green). **b**, Box plots of Ca^{2+} transient amplitudes ($\Delta F/F_0$) of L2/3 S1 neurons during stimulus presentation (s_1 or s_2) for correct trials from WT control and lOFC silenced animals (50, 50 neurons the LE, RE phase respectively from 4 wildtype mice and 42, 32 neurons from 3 lOFC silenced mice) during expert phases (LE, RE). Chemogenetic silencing of lOFC during the reversal phase was achieved by virally induced expression of hM4di in excitatory pyramidal neurons and daily systemic CNO injection prior to each session. Statistical differences were tested with two-tailed, independent samples T-tests with Bonferroni corrections. **c**, Box plots of the summed firing rates of the simulated sensory neurons, $\sum_i z_i$, in response to stimulus 1 (s_1) and stimulus 2 (s_2) for intact and silenced lOFC. **d**, Functional schematic of panel a, showing the mirror-symmetry between the prediction of reward and the prediction of confidence loss. The contribution of a sensory neuron to the prediction of the future reward determines its neuron-specific value (blue). The context-prediction error (green) is a signal anticipating a confidence loss, which hence counteracts the reward-guided sensory amplification ahead of the outcome. **e**, Final performance (after 600 trials, RE) with and without lOFC for the original and for the reversed rule. The pie charts show the fraction of agents that reached expert performance. 64 pseudorandom initializations were simulated for each condition. **f**, Cumulative distributions of simulated agents that reached expert performance (85% correct) along the progression of trials; either with (left) or without (right) lOFC-mediated apical inhibition via SST. **g**, Performance traces of the agents with and without lOFC during reversal learning. Bottom: box plots show the distributions of the number of trials necessary to become expert for the original or the reversed rule.

end, the synaptic weights exciting formerly amplified neurons had to be weakened; and different synapses had to be strengthened. An alternative strategy is to modify the sensory amplification by switching the context representation while maintaining the synaptic weights that transmit the top-down excitation [13] (Fig. 4f). It has been shown that sensory amplification is context-dependent [5, 21, 32, 33] and that animals can develop dedicated task-representing neurons [34]. Furthermore, activity in the human lOFC has been suggested to represent a shift in context representation [27]. We, therefore, explored whether this principle would yield computational benefits. For that, we simulated agents with two possible context representations, each consisting of a two-dimensional one-hot vector and projecting to the apical dendrites of the sensory neurons. The lOFC activity encoding the context-prediction error was then used to trigger a switch between context representations (Fig. 4f). The capacity to represent different contexts and to switch between these representations significantly improved behavioral adaptation following multiple rule reversals (Fig. 4g). Indeed, when a rule had already been experienced previously, not only were the stimulus-to-action projections already learned (as in the single-context model), but also the top-down excitation weights onto the sensory neurons (Extended Data Fig. S8a). While we implemented our two-context model to also include lOFC to sensory cortex SST interneuron feedback, this was somewhat redundant as the context-representation switch prevented the inadequate sensory amplification from the prior rule. Nonetheless, SST interneuron activation could provide a necessary inhibition in case the new context-representation was established and had yet to be formed.

VIP-mediated disinhibition reinforces anticipatory sensory signals

We showed how top-down signals such as the context-prediction error could improve sensory processing to accommodate different behavioral contexts. Nonetheless, while we provided a mathematical framework, how the lOFC learns to generate the context-prediction error signal during the stimulus window is still unclear. In general, through the accumulation of experience across trials, anticipatory responses build up ahead in time from the reinforcers they learn to predict. While a variety of such predictive signals can be observed in the cortex [35, 36], little is known about the underlying microcircuits and neural mechanisms that generate them. Accumulating evidence shows that VIP-mediated

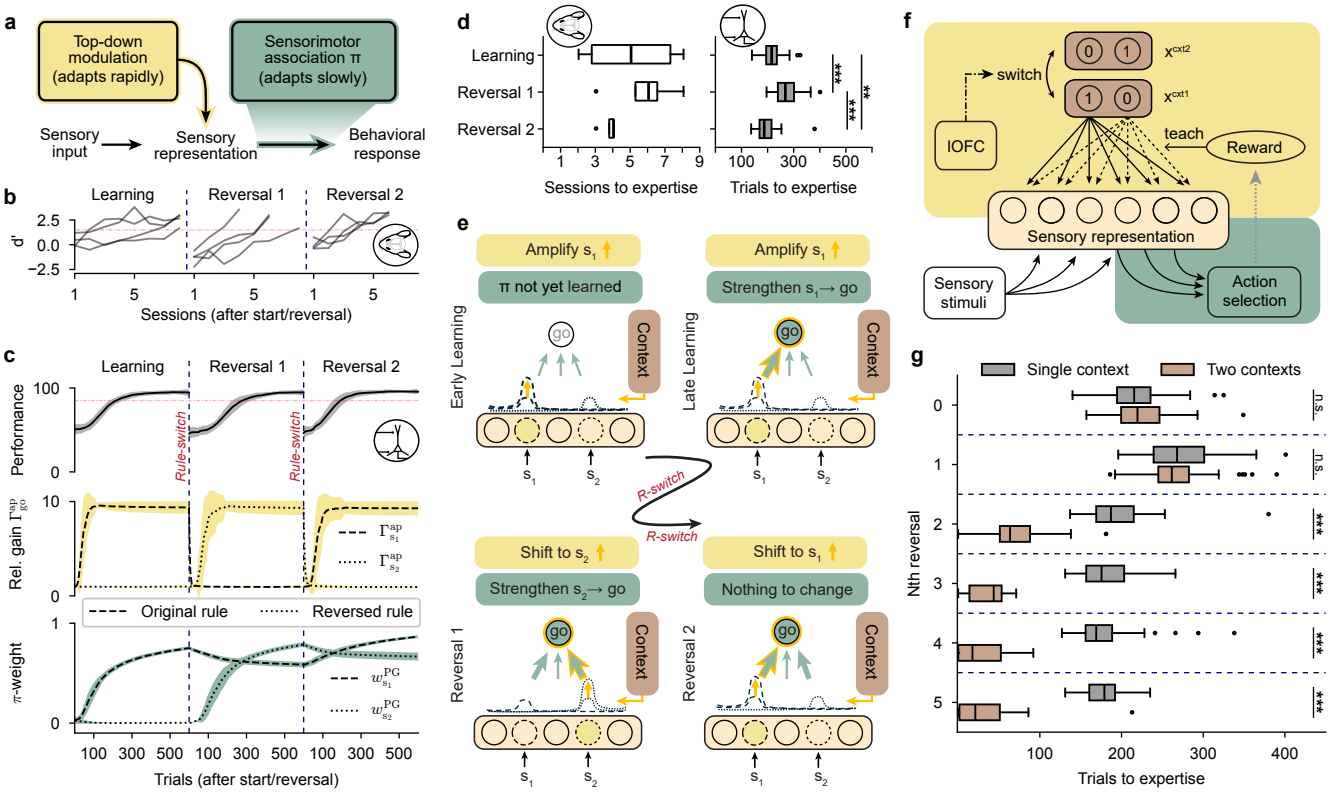

Fig. 4 The context-prediction error reduces adaptation time for multiple reversals. a, The two main components involved in behavioral learning: rapidly adapting top-down modulation of the sensory representation depending on the context (yellow), and slower adapting sensorimotor associations π (green). **b**, Performance traces of 4 mice during two consecutive rule reversals, where the second one reverts back to the original rule. **c**, Top: Average performance of the agents for two sequential reversals. Middle: Evolution of the relative gain, see equation (12), of the go-stimulus-selective neurons, which corresponds to the s_1-selective neuron when the original rule applies (dashed) and to the s_2-selective neuron for the reversed rule (dotted). After each reversal, the top-down inhibition rapidly erases the relative gain. Next, the π-weight represents the synaptic strength of the weight in the policy network connecting from the go-stimulus-selective neuron to the go-action neuron. After potentiating the synaptic weight between the s_1 neuron and the go action during Learning, it is maintained throughout Reversal 1, while the synaptic weight between the s_2 neuron and the go action is strengthened. **d**, Number of sessions needed for each of the 4 mice to reach expertise for Learning, Reversal 1, and Reversal 2 (left) and the same in trials for the model (right). **e**, Qualitative illustration of the steps that led to faster adaptation after the second reversal. Early learning: amplification of the s_1-selective sensory neuron. Learning: increasing the synaptic weight from the s_1-selective neuron to the go-action neuron. Reversal 1: top-down sensory amplification is shifted from the s_1 to the s_2-selective sensory neuron, and the weight from the s_2-neuron to the go-action neuron is strengthened. Reversal 2: synaptic weights from the s_1-selective neuron to the go-action neuron remain strengthened, only the sensory modulation needs adjusting. **f**, Model in which two different contexts can be represented and where the context representation x^{aP} can be switched by lOFC activity (context-prediction error). Allowing the sensory modulation to be adapted by changing x^{aP} removes the need to re-learn the top-down modulation weights in the case when a previously learned context reappears. **g**, Number of trials necessary to reach expert performance with a single context representation (grey) and with two context representations (brown) for five sequential rule reversals. Statistical differences were tested with two-tailed, independent samples T-tests with Bonferroni correction.

disinhibition [37–39] may play a critical role in gating associative learning in the olfactory [40], auditory [41], visual [42, 43] and somatosensory cortex [44], and hence could be part of a cortex-wide mechanism [45]. Hence, VIP interneurons could be responsible for signaling a predictive target for pyramidal neurons. In different cortical regions they are recruited by reinforcers such as reward and punishment [46]. Applied to our model of lOFC, we therefore suggest that lOFC VIP interneurons could convey the confidence loss (CL) produced by unexpected outcomes (Fig. 3d). In consequence the inhibition of lOFC VIP neurons could be modeled by setting CL = 0 in our model.

We pharmacogenetically silenced the activity of VIP interneurons in VIP-cre mice by injecting Cre-dependent hM4Di and administering intra-peritoneal CNO during reversal learning. When inhibiting lOFC VIP interneurons after reversing stimulus-outcome association, both in mice and the computational model, we found that silencing VIP interneurons in lOFC had a critical effect in significantly delaying reversal learning (Fig. 5a). The model suggests that the degraded reversal learning can be attributed to a missing top-down inhibition from lOFC through local SST interneuron interaction in S1. In the control condition, the model uses lOFC activity to drive sensory SST neurons, as experimentally observed in V1 [20]. If VIP interneurons in lOFC are silenced, the confidence loss triggered during the outcome window is prevented to reinforce the anticipatory context-prediction error during the stimulus window (Fig. 2c). As a consequence, the previously learned apical values in S1 pyramidal neurons are not suppressed by top-down inhibition upon the rule-reversal (Fig. 5b,c). In S1, the previous reward-predictive sensory neurons remain enhanced

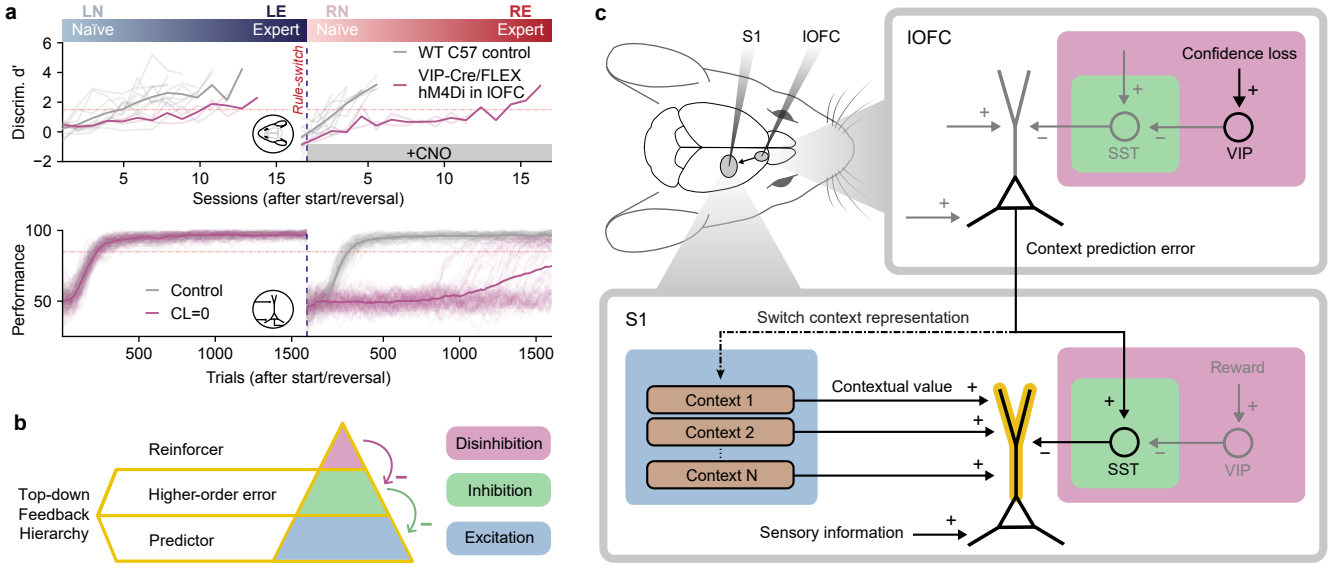

Fig. 5 VIP interneurons as local reinforcers for learning a hierarchy of anticipatory signals. a, Top: Performance of mice during reversal learning with (grey) and without (purple) chemogenic silencing of their lOFC VIP neurons (10 and 2 mice respectively). Bottom: Simulated performances with the confidence loss (CL) fixed to zero (grey) compared to the undisturbed control case (purple). **b**, Top-down feedback hierarchy onto pyramidal neurons, where inhibition overrules excitation, and disinhibition overrules inhibition. At the lowest level (blue), the top-down excitation targetting the apical dendrites of pyramidal neurons represents a context-dependent anticipatory predictor (in S1 of reward, in lOFC of a confidence loss). At the middle level (green), the SST-mediated top-down inhibition counteracts the lower level excitation by a higher-order error (in S1 the context-prediction error counteracting the neuron-specific apical value). At the top level (purple), the top-down disinhibition represent a reinforcer conveyed by VIP interneurons (in S1 an action-induced reward signal, in lOFC the confidence loss). **c**, Microcircuits implementing the top-down feedback hierarchy from b within S1 and lOFC, with apical excitation (blue) inhibited by an error-signal (green), and a new anticipatory context signal (e.g. Context 2 in S1) being reinforced through disinhibition.

and learning is disturbed, because the action selection network has to compensate for the incorrectly modulated sensory representation (see also Fig. 4d).

Discussion

We presented dual calcium imaging in S1 and lOFC of mice during a stimulus-response association task while switching the association rule. During learning of a first rule, the sensory response in S1 to the reward-predictive stimulus was enhanced. After reversing the rule, the stimulus enhancement to the previously reward-predictive stimulus is undone, while the response to the reward-predictive stimulus in the second rule is now getting enhanced. We experimentally identified lOFC as the seat to signal the rule-switch and trigger the reorganization of the sensory representation in S1. A theory is presented that explains the top-down induced reorganization of S1 based on the notion of a context-prediction error.

During learning, the confidence in the predicted outcome increases. An unexpected rule-switch results in a confidence loss at the moment, when the previously correct outcome prediction turns out to be wrong. We showed that such a confidence loss is signaled by lOFC after the rule switch, when the expected reward during the outcome window is missed. Crucially, in trials following the rule-switch, lOFC advanced its firing and learned to signal a context-prediction error already during the stimulus window, being predictive for the confidence loss confirmed in the following outcome window. In S1, the sensory response to the reward-predicting stimulus is enhanced. The stimulus enhancement is neuron-specific and proportional to the contribution of the neuron to the reward prediction. With the signaling of the context-prediction error by lOFC after the rule-switch, the response enhancement in S1 is undone via apical inhibition through local SST interneurons. During the subsequent relearning, the sensory response to the newly reward-predictive stimulus becomes enhanced through a reorganization of the top-down input to S1. Blocking experiments suggest that VIP interneurons in lOFC signal the confidence loss to the lOFC pyramidal neurons, providing them the teaching for learning the anticipatory context-prediction error. This context-prediction error triggers the reorganization of the top-down induced gain modulation in S1.

The behavior-dependent reorganization of sensory representation during learning is well recognized [47–49], and top-down feedback has been established as a fundamental component of perceptual learning [8, 10, 11, 13, 15, 16, 25]. Our work revealed how early regions of sensory processing can be modulated by lOFC to reshape sensory computation

depending on task context. The importance of the lOFC in reversal learning and contextual response-outcome associations has been well documented in mice [18], rats [50], non-human primates [51], and humans [19, 52]. We recently reported that similar neural circuits involving lOFC could be operational using analogous cognitive strategies in distinct species [53]. While many different roles have been proposed for the lOFC to support cognitive flexibility [54, 55], its importance has been highlighted for state representation learning [56–58], and in particular for representing confidence [59, 60] and context switching [27]. We demonstrated how the lOFC, by signaling a context-prediction error, can reshape sensory processing after a rule switch.

The hierarchy of reward- and context-prediction errors can be viewed as a mechanism to reduce different sources of uncertainties, in line with other uncertainty minimization theories [61]. Selective amplification prevents irrelevant information to disturb the learning process, which reduces the expected uncertainty [62] as long as the context remains the same. While uncertainty estimates have been suggested for global learning-rate modulation [63], regulating selective top-down amplification provides a mechanism for a neuron-specific learning-rate modulation (see Supplementary Equations 1). A rule reversal elicits uncertainty about the current context, which can be termed unexpected uncertainty (see Supplementary Equations 3), and should reduce reliance on top-down priors [62]. SST interneurons have been shown to drive a shift away from top-down influences to favor bottom-up information streams [31]. Lateral OFC projections to SST interneurons in S1 provide a direct mechanism to inhibit selective top-down amplification and unbias the sensory representation in case of contextual uncertainty.

VIP disinhibition represents a local reinforcement signal that varies depending on the cortical region [44, 46, 64]. In the basolateral amygdala, VIP interneurons respond strongly to punishments and are necessary to develop enhanced responses to conditioned stimuli during fear learning [65]. The only ionotropic serotonin receptors (5-HT3R) expressed in the cortex are found in interneurons that do not express PV or SST [66] and co-express VIP and CCK [67]. Multiple studies have found that serotonin in the OFC plays a crucial role for reversal learning [68–70]. As serotonin has been associated with sensory uncertainty [71], unsigned reward-prediction error [72], surprise [73], expected and unexpected uncertainty [63], it makes an interesting suspect to relay confidence loss to lOFC VIP interneurons (Fig. 5c).

Our computational model provides a structure to makes prediction that can be experimentally tested. For example, is lOFC necessary only for the first rule reversal and not the subsequent ones after context representations have been consolidated? If the function of lOFC is purely based on SST-mediated inhibitory feedback (Fig. 3a), the model predicts that a strong lOFC response will be required following each reversal. In the case where lOFC directly changes the context representation (Fig. 4f), the model predicts that the lOFC response upon a rule reversal would decrease for subsequent reversals (Extended Data Fig. S8b).

The richness with which the brain modulates sensory representations through first- and second-order reward-prediction errors could also have applications in machine learning. Dendritic nonlinearities were suggested to increase the expressive power of neural networks [74] and context-dependent dendritic gain modulation was proposed as a biologically plausible mechanism for multi-task learning and as a means to reduce catastrophic forgetting in continual learning [13, 34, 75, 76]. However, these models lacked an explanation for how the brain can recognize a change of context, and signal the context change before the decision making, during sensory perception. Prospective errors were suggested to project to apical dendrites of pyramidal neurons and prospectively correct the activity of deep neurons in the cortical hierarchy [77–80]. The presented data and theory on context-prediction errors is in line with such previous results and offers new insights that may inspire future developments in both computational neuroscience and artificial intelligence [3, 81].

Methods

Gain modulation in sensory representations

The sensory representation $\boldsymbol{z}$ consists of N pyramidal neurons $i \in \{1, ..., N\}$ with two compartments. The basal compartment received bottom-up sensory inputs with fixed selectivity producing activations $\boldsymbol{x}^{\text{bas}}$ and the apical compartment dictated the multiplicative gain $\boldsymbol{g}^{\text{ap}}$ of the firing rates,

$$z_i = g_i^{\text{ap}} \lfloor x_i^{\text{bas}} \rfloor \,, \tag{1}$$

where $\lfloor \cdot \rfloor$ stands for the rectifier function $\lfloor y \rfloor = \max(0, y)$. An apical compartment receives $N_{\text{cxt}} = 2$ context-dependent excitatory inputs x_j^{cxt}, weighted by the synaptic strength w_j^{ap}. It also receives global inhibition r^{SST} from a population of SST interneurons, see equation (10). These top-down apical afferents determine the context-specific apical gain g_i^{ap} within the range $[1, g_{\max}]$,

$$g_i^{\text{ap}} = \min \left\{ 1 + \left\lfloor \sum_{j=1}^{N_{\text{cxt}}} w_{ij}^{\text{ap}} x_j^{\text{cxt}} - r^{\text{SST}} \right\rfloor \,,\, g_{\max} \right\} \,. \tag{2}$$

The *salience* of a sensory stimulus is proportional to the gain g_i^{ap} of its sensory neurons. The gain is amplified through the apical excitation $\sum_{j=1}^{N_{\text{cxt}}} w_{ij}^{\text{ap}} x_j^{\text{cxt}}$ that originates in the *context representation* $\boldsymbol{x}^{\text{cxt}}$. The apical excitation can be cancelled by the SST interneuron activity r^{SST}, or altered by changing $\boldsymbol{x}^{\text{cxt}}$ (as in Fig. 4f).

Apical amplification based on neuron-specific value

The apical weights $\boldsymbol{w}^{ap}$ are learned to amplify sensory neurons that are predictive for reward. Each neuron is assigned a neuron-specific value v_i, encoding how much this neuron contributes to the estimate of the (global) value $V(z) = \mathbb{E}[R|\boldsymbol{z}]$, i.e. to the expected future reward conditioned the current state representation. The value estimator $\hat{V}(\boldsymbol{z})$ is calculated based on the sensory representation $\boldsymbol{z}$ as

$$\hat{V}(\boldsymbol{z}) = \sum_i w_i^V z_i \ . \tag{3}$$

The weights w_i^V are adapted to reduce the reward-prediction error $\text{RPE} = R - \hat{V}(\boldsymbol{z})$, constrained by a L1 regularization,

$$\Delta w_i^V = \eta^V \left(R - \hat{V}(\boldsymbol{z})\right) z_i - \lambda^V \operatorname{sgn}(w_i^V) \ . \tag{4}$$

This weight update reduces the loss $L^V = \mathbb{E}[\text{RPE}^2] + \lambda^V \|\boldsymbol{w}^V\|_1$, where $\|.\|_1$ refers to the L1-norm. The neuron-specific value v_i encoding how much a neuron i predicts an upcoming reward is defined by its non-negative contribution to the value estimation $\hat{V}$, i.e. $v_i = \lfloor w_i^V z_i \rfloor$. A key ingredient of the model is that the sensory representation is sharpened by the top-down input, such that the task-relevant sensory neurons are amplified compared to task-irrelevant neurons. The criterion for task-relevance is set by the comparison between the value v_i of a neuron and the activity r^{SST} of the SST interneurons. If $v_i > r^{\text{SST}}$, the apical excitation is enhanced. If $v_i < r^{\text{SST}}$, the neuron i is not sufficiently relevant and its apical excitation is reduced. This yields to the learning rule for the excitatory apical synapses

$$\Delta w_{ij}^{\text{ap}} = \eta^{\text{ap}} \left(v_i - r^{\text{SST}}\right) x_j^{\text{ap}} \ , \tag{5}$$

where η^{ap} represents the learning rate and x_j^{ap} is the presynaptic activity, see equation (2). Asymptotically, this plasticity rule leads to a binary distribution of the apical weights. Active synapses projecting to a highly task-relevant sensory neuron become potentiated and converge towards $w_{\text{max}}^{\text{ap}} = 1$. Synapses projecting to a task-irrelevant neurons become depressed and converge towards $w_{\text{min}}^{\text{ap}} = 0$ (see Supplementary Equations 2). Taken together, the RPE informs the value estimate and hence the gain amplification of individual L2/3 sensory neurons.

Estimating the confidence of the reward prediction

In order for the brain to notice a change of context, it first needs a notion of understanding the current context at hand. The action-value estimate $\hat{Q}(a, \boldsymbol{z})$ approximating $\mathbb{E}[R|a, \boldsymbol{z}]$ — while also supporting action selection (see Supplementary Methods 2) — encodes the current understanding of the agent about the task. The action-value estimate $\hat{Q}(a, \boldsymbol{z})$ is computed from an outcome prediction network that implicitly represents the the context, i.e. the task rule, through neurons becoming selective for stimulus-action-outcome triples (see Supplementary Methods 1).

The variance unexplained (VU) by the model can be measured by the expected squared reward-prediction error $\sigma^2 = \mathbb{E}[(R - \hat{Q})^2]$, see Supplementary Equations 4. The variance unexplained, σ^2, jointly measures the bias in the reward prediction, its distraction by task-irrelevant stimuli, and the intrinsic variability of R (see Supplementary Equations 3). We assume that the brain approximates σ^2 with an internal estimate that we denote as *prediction uncertainty* $\hat{\sigma}^2$ [82]. The prediction uncertainty $\hat{\sigma}^2$ is updated based on the difference between the current squared reward-prediction error at the current time step and the uncertainty estimate in the previous trial,

$$\Delta\hat{\sigma}^2 = \eta_\sigma \left((R - \hat{Q})^2 - \hat{\sigma}^2\right) \ , \tag{6}$$

with a small learning rate η_σ. If $\hat{\sigma}^2$ is small, the animal can be confident in its model of the task and the reward prediction. We therefore defined the *prediction confidence* c as estimate of the squared *fraction of explained variance* (FVE), or also squared coefficient of determination (see Supplementary Equations 4),

$$c = \left\lfloor 1 - \frac{\hat{\sigma}^2}{\sigma^2_{\text{max}}} \right\rfloor^2 \ . \tag{7}$$

Here, $\hat{\sigma}^2/\sigma^2_{\text{max}}$ represents the fraction of variance unexplained (FVU). The maximal variance $\sigma^2_{\text{max}} = 0.5$ was computed for a uniform random policy (see Supplementary Equations 5). The prediction confidence c is restricted to values between 0 and 1. For more details see Supplementary Equations 4.

Confidence loss from unexpected outcomes

The prediction uncertainty $\hat{\sigma}^2$ can also be understood as a form of expected uncertainty given the internal model of the task [62, 83]. The prediction confidence, correspondingly, can be seen as model confidence. When an outcome occurs that lies further away from the prediction than would be expected, it can elicit unexpected uncertainty [63]. Translated to our framework, if $(R-\hat{Q})^2 - \hat{\sigma}^2 > 0$, the brain signals a higher-order error compared to the first order prediction error $R-\hat{Q}$. As the rat OFC has been shown to be involved in confidence computation [60], we suggest that the *confidence loss* (CL) could also be signaled to lOFC. In our model, CL is defined as decrease in model confidence induced by the change $\Delta\hat{\sigma}^2$ given in equation (6). The confidence loss can be expressed as a function of $\Delta\hat{\sigma}^2$,

$$\mathrm{CL} = \lfloor -\Delta c \rfloor \approx \frac{2\sqrt{c}}{\sigma^2_{\max}} \lfloor (R-\hat{Q})^2 - \hat{\sigma}^2 \rfloor = \frac{2\sqrt{c}}{\sigma^2_{\max}} \lfloor \Delta\hat{\sigma}^2 \rfloor \,, \tag{8}$$

see Supplementary Equations 6. Hence, the CL can be viewed as a confidence-weighted 2nd-order error of the reward prediction. A strong confidence loss is contingent on the animal being confident in its model of the environment ($c > 0$).

Context-prediction error in lOFC anticipates the confidence loss

We denote the anticipation of the CL at stimulus time the *context-prediction error* (CPE). While the confidence loss CL is elicited when an unexpected outcome is observed during the reward presentation window t_R, the contextual uncertainty can be represented in the next trial already during the stimulus window t_S. The CPE is learned by minimizing the squared difference between CL and CPE, $\frac{1}{2}(\mathrm{CL}-\mathrm{CPE})^2$. Minimizing this squared difference online by gradient descent leads to an update of the CPE by

$$\Delta\mathrm{CPE} = \eta_\mathrm{lOFC}\ (\mathrm{CL}-\mathrm{CPE})\ , \tag{9}$$

with learning rate η_lOFC. In the simulations, we model the lOFC activity during the stimulus window as context-prediction error, $r^\mathrm{lOFC} = \mathrm{CPE}$.

Context-prediction error inhibits apical dendrites in S1 via SST interneurons

The SST interneurons in S1 are driven by the context-prediction error encoded in lOFC. They modulate the sensory processing in two ways. First, they directly reduce the gain of the sensory representation neurons in S1 by inhibiting the apical dendrite activity, see equation (2). Second, they represent a task-relevance threshold that the neuron-specific value v_i has to reach in order to potentiate excitatory apical synapses, see equation (5) and Supplementary Equations 2. The firing rate of the SST population in S1 at the time of stimulus presentation t_S was modeled as

$$r^\mathrm{SST}(t_\mathrm{S}) = w^\mathrm{SST}\ \left\lfloor r^\mathrm{lOFC} - r_0^\mathrm{VIP} \right\rfloor + r_0^\mathrm{SST}\ , \tag{10}$$

where r_0^SST, w^SST and r_0^VIP are constants.

lOFC triggers switch of context representation

In most simulations, the gain amplification conveyed by excitatory apical synapses $\boldsymbol{w}^\mathrm{ap}$ originated from a single, fixed context representation $\boldsymbol{x}^\mathrm{cxt} = (1, 0)$. For this single-context model of the brain, the apical gain could only be adapted by changing the synaptic weights $\boldsymbol{w}^\mathrm{ap}$ or by tuning the inhibitory contribution from SST interneurons. Alternatively, different contexts representations $\boldsymbol{x}^\mathrm{cxt}$ can be considered, and the gain can be modified by switching the context. To show this, we implemented context 1 by $\boldsymbol{x}^\mathrm{cxt1} = (1, 0)$ and context 2 by $\boldsymbol{x}^\mathrm{cxt2} = (0, 1)$. The switch between both contexts was triggered when the lOFC activity reached a threshold activity θ_switch and the model confidence c was above a threshold c_switch.

Confidence-weighted action selection

During reversal learning, the policy depended on the sensory representation $\boldsymbol{z}$ and on the model-confidence c. The probability of selecting an action $\pi(a|\boldsymbol{z}, c)$ consisted of a combination of two separate action selection policies: one explorative policy π^Q based on action-values and one exploitative policy π^PG based on policy gradient (Extended Data Fig. S4a)

$$\pi(a|\boldsymbol{z}, c) = c \cdot \pi^\mathrm{PG}(a|\boldsymbol{z}) + (1-c) \cdot \pi^Q(a|\boldsymbol{z})\ . \tag{11}$$

Hence, at low confidence c — in case of high uncertainty $\hat{\sigma}^2$ — action selection was mainly driven by the explorative policy π^Q. At high confidence c — in case of low uncertainty — the action selection was dominated by the exploitative policy π^PG. For more details on the action selection (see Supplementary Methods 2, 3 and Extended Data Table 1).

Simulated go/no-go task

Learning of the go/no-go sensory discrimination task was simulated for 600 trials (Fig. 1b). After that, the simulation was either terminated or the stimulus-reward contingency was reversed. Each trial began by randomly sampling the N-dimensional stimulus pattern $\boldsymbol{x}^{\text{bas}}$ characterizing the basal input to the N sensory pyramidal neurons. In all simulations, one sensory representation neuron was exclusively selective to stimulus 1 (s_1), which means that its basal activation $\boldsymbol{x}^{\text{bas}}_{s_1}$ was 1 in case the stimulus 1 was presented and 0 otherwise. Analogously, one neuron was exclusively selective to stimulus 2 (s_2). The remaining $N_d = N-2$ 'distractor' neurons received random basal inputs independently sampled from binary noise $\xi_i \in \{0, 1\}$ with probability 0.5 for each trial (see Fig. 1c). In other words, x_i^{bas} was either 1 or 0 (except for Fig. 1e, 1f, Extended Data Fig. S4d and S4e, where the s_1- and s_2-dependent response strengths were varied). Unless specified otherwise, we used $N_d = 128$ and $N = 130$.

The bottom-up activations $\boldsymbol{x}^{\text{bas}}$ were processed to compute the sensory representation $\boldsymbol{z}$, see equation (1). Based on $\boldsymbol{z}$, the policy π then selected between two actions: go (a_{GO}) or no-go (a_{NG}). Depending on the stimulus and action combination, an outcome with reward $R \in \{R_{\text{punish}}, 0, R_{\text{reward}}\}$ was fed back to the agent (Extended Data Fig. 1b).

Simulation parameters

All simulations were run with 64 different pseudorandom seeds, which were all used to compute the means and other statistics displayed in the figures. When applicable, the parameters $\eta^V, \eta^\pi, \eta^Q, r_0^{\text{VIP}}$ and w^{SST} were set by a grid search algorithm (Extended Data Algorithm 2) to either maximize the performance or minimize the number of trials necessary to reach expert performance (Extended Data Table 3). All other parameters (λ^V, r_0^{SST}, T^Q, η_c, $g_{\max}$, $\sigma^2_{\max}$, R_{reward}, R_{punish}) were predefined constants (Extended Data Table 2). A detailed description of the parameter settings is included in Supplementary Methods 3 and 4.

Model evaluation

In order to evaluate the computational model simulations and compare them to experimental observations, we employed different metrics:

Performance. The learning progress was quantified by tracking the performance during the simulation. The performance is computed as a moving average of the number of correct actions taken within a window of 100 trials. This was implemented with a convolution using a reflection padding of 50 trials at the boundaries (beginning / end of training or before / after reversals). The performance threshold to achieve expertise was set at 85%.

Relative gain. The relative gain of the sensory neuron i used in Fig. 4a quantifies how the gain g_i^{ap} of a pyramidal neuron i compares to the gain of the other neurons. It is defined as the gain divided by the average gain across all other sensory neurons,

$$\Gamma_i^{\text{ap}} = \frac{g_i^{\text{ap}}}{\frac{1}{N}\sum_{j\neq i}^{N} g_j^{\text{ap}}} \,. \tag{12}$$

The relative gain for the go-stimulus selective neuron $\Gamma_{\text{go}}^{\text{ap}}$ will be taken for a different sensory neuron depending on the current task rule (as in Extended data Fig. S8a), unless it is specified for an explicit rule (as in Fig. 4)c).

π-weight. The strength of synaptic weight $w_{\text{GO}}^{\text{PG}}$ originates from the go-stimulus neuron (the s_1-selective neuron during the original rule and the s_2-selective neuron in case of the reversed rule) and targets the go action neuron in the policy network π^{PG}. The π-weight measures $w_{\text{GO}}^{\text{PG}}$ and characterizes the sensorimotor association that has to be learned for the go/no-go sensory discrimination task.

Signal-to-noise ratio. The signal-to-noise ratio (SNR) used in Fig. S3 is defined as the ratio between the variance of a stimulus encoding neuron and of the distractor neurons. Since the activity z_s of a go/no-go stimulus encoding neuron corresponds to a Bernoulli variable (with probability $p = 0.5$) amplified by a gain g_s and the sum of distractor neuron activities $\sum_{i=1}^{N_d} z_i$) corresponds to a binomial distribution (also with probability $p = 0.5$ and in this example without gain modulation), we find that

$$\text{SNR} = \frac{\text{Var}(z_s)}{\text{Var}(\sum_{i=1}^{N_d} z_i)} = \frac{p(1-p)g_s^2}{p(1-p)N_d} = \frac{g_s^2}{N_d} \,. \tag{13}$$

Animals

C57BL/6, male mice aged between 8 and 16 weeks were used in our experiments. All experimental procedures were conducted in accordance with the requirements of the United Kingdom Animals (Scientific Procedures) Act 1986 and the Federal Veterinary Office of Switzerland, approved by the Cantonal Veterinary Office in Zürich. Experiments were conducted under the authority of Home Office Project License PABAD450E and personal license numbers 285/2014 and 234/2018. Mice were housed at 24°C in a 12-h reverse dark-light cycle (07h00 to 19h00). At the end of the experiment, the mice were deeply anaesthetized, transcardially perfused and euthanized by exposure to CO2 in their home cage.

Water scheduling

During behavioural training, mice stayed on a 5/2 days water restriction paradigm. They were weighed daily and kept at 85-90% of the previously determined baseline weight. In case animals did not receive enough reward during training sessions (40ml/kg/day), they were given water after training. Water was always provided ad libitum during weekends.

Handling and habituation

To familiarize the animals with the experimenter, handling was started before surgeries and continued during the recovery period. Habituation to head restraint in progressively lengthened sessions until they tolerated at least 15 minutes of head restraint.

Cranial window and cannula implantation

Mice were implanted with a cranial window (S1) or cannula (lOFC) under controlled anaesthesia [18]. For the cranial window preparation, a 4 mm cover glass was placed inside the craniotomy and the edges were sealed with dental cement. For imaging deeper structures, the superficial brain tissue was aspirated, a cannula with a glass bottom placed on top of the target area and the edges sealed with dental cement. A head post was attached using dental cement, and the skin was reattached to the implant.

Sensory learning and reversal learning task

We trained mice on a tactile Go/No-Go reversal learning task [18]. Animals were presented with two sandpaper textures (P100, coarse, s_1; P1200, fine, s_2) and they needed to learn to discriminate to either earn a sucrose water reward ('Go'-texture, P100, 'hit') or withhold a response to avoid punishment ('No-Go'-texture, P1200, 'CR', correct rejection). Responding towards the 'No-Go'-texture resulted in experiencing white noise ('FA', false alarm). After animals displayed stable expert performance (d'>1.5 in three consecutive sessions; d'$= Z(\text{hit}/(\text{hit}+\text{miss})) - Z(\text{FA}/(\text{FA}+\text{CR}))$, with $Z(p), p \in [0,1]$, being the inverse of the cumulative Gaussian distribution), stimulus-outcome associations were reversed (P100, 'No-Go'-texture; P1200, 'Go'-texture) and mice needed to adapt their responses accordingly. The task was amended into a serial reversal learning task by introducing a second rule-switch after animals displayed stable expert performance during initial reversal learning (d'>1.5 in three consecutive sessions), reinstating initial stimulus-outcome (P100 as 'Go'-texture, P1200 as 'No-Go'-texture).

Two-photon Ca^{2+} imaging

Neurons were imaged using a custom-built two-photon microscope equipped with a Ti: Sapphire laser system (approximately 100 femtosecond laser pulses; Mai Tai HP, Newport Spectra Physics), a water-immersion 16X Olympus objective (340LUMPlanFl/IR, 0.8 numerical aperture, NA), galvanometric scan mirrors (model 6210; Cambridge Technology) and a Pockels Cell (Conoptics) for laser intensity modulation. Virally expressed GCaMP6f in L2/3 neurons in S1 were imaged at 940 nm. Neuronal activity was recorded at a 15 Hz frame rate (796 x 512 pixels) for 10 seconds each trial by detecting fluorescence changes using a photomultiplier tube (Hamamatsu). Imaging was intermitted during the 3s ITI.

Ca^{2+} imaging analysis

Ca^{2+} imaging data of each session was processed using the Python-based Suite2p processing pipeline ([84]). The recordings were motion-corrected using non-rigid registration. Cells were detected with automated ROI detection followed by manual curation. Raw fluorescence time courses were extracted as each ROI's (non-weighted) mean pixel value. In addition, a neuropil trace was computed for each ROI to generate neuropil-corrected calcium traces F(t). The fluorescence change during each trial epoch (stimulus-evoked, reward-responsive responses) was then calculated by subtracting the baseline fluorescence F_0 (mean fluorescence value of the first 1.5 s) from the corrected traces F and dividing by F_0; $\Delta F/F_0 = (F - F_0)/F_0$.

Chemogenic silencing

Inhibitory DREADDs (hM4Di) were either expressed in excitatory (CaMKIIα-hM4D(Gi)-mCherry) or GABAergic VIP-interneurons (AAV1/2-hSyn1-dlox-hM4D(Gi)-mCherry(rev)-dlox, VIP-cre mouse line) in layer 2/3 of the lOFC through virally mediated transfection. hM4Di was activated via intraperitoneal (i. p.) injection of the ligand clozapine N-oxide (CNO dihydrochloride, 1–5 mg/kg, Tocris Bioscience, Catalog number 4936) 25-30 minutes before each training session.

Data availability

The experimental data that support the finding of this study are available upon reasonable request from the corresponding author. The simulated data can be reproduced from the code provided.

Code availability

The code for the simulations and visualisations is provided on https://github.com/unibe-cns/TopDownOFC.

References

[1] Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).

[2] Niv, Y. Reinforcement learning in the brain. *Journal of Mathematical Psychology* **53**, 139–154 (2009).

[3] Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).

[4] Treue, S. & Trujillo, J. C. M. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575–579 (1999).

[5] Atiani, S., Elhilali, M., David, S. V., Fritz, J. B. & Shamma, S. A. Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron* **61**, 467–480 (2009).

[6] Makino, H., Hwang, E. J., Hedrick, N. G. & Komiyama, T. Circuit mechanisms of sensorimotor learning. *Neuron* **92**, 705–721 (2016).

[7] Salinas, E. & Thier, P. Gain modulation: a major computational principle of the central nervous system. *Neuron* **27**, 15–21 (2000).

[8] Li, W., Piëch, V. & Gilbert, C. D. Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience* **7**, 651–657 (2004).

[9] Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E. & Golestani, N. Cortical encoding of speech enhances task-relevant acoustic information. *Nature human behaviour* **3**, 974–987 (2019).

[10] Yang, Y., Shen, H. & Kwon, S. E. Learning-induced reorganization of neuronal subnetworks in the primary sensory cortex. *bioRxiv* 2023–02 (2023).

[11] Huang, S., Wu, S. J., Sansone, G., Ibrahim, L. A. & Fishell, G. Layer 1 neocortex: Gating and integrating multidimensional signals. *Neuron* (2023).

[12] Larkum, M. E., Senn, W. & Lüscher, H.-R. Top-down dendritic input increases the gain of layer 5 pyramidal neurons. *Cerebral cortex* **14**, 1059–1070 (2004).

[13] Wybo, W. A. *et al.* NMDA-driven dendritic modulation enables multitask representation learning in hierarchical sensory processing pathways. *Proceedings of the National Academy of Sciences* **120**, e2300558120 (2023).

[14] Benezra, S. E., Patel, K. B., Campos, C. P., Hillman, E. M. & Bruno, R. M. Learning enhances behaviorally relevant representations in apical dendrites. *bioRxiv* 2021–11 (2021).

[15] Schoenfeld, G. *et al.* Unsigned temporal difference errors in cortical l5 dendrites during learning (2024).

[16] Schäfer, R., Vasilaki, E. & Senn, W. Perceptual learning via modification of cortical top-down signals. *PLoS computational biology* **3**, e165 (2007).

[17] Marvan, T., Polák, M., Bachmann, T. & Phillips, W. A. Apical amplification—a cellular mechanism of conscious perception? *Neuroscience of consciousness* **2021**, niab036 (2021).

[18] Banerjee, A. *et al.* Value-guided remapping of sensory cortex by lateral orbitofrontal cortex. *Nature* **585**, 245–250 (2020).

[19] Wang, B. A., Veismann, M., Banerjee, A. & Pleger, B. Human orbitofrontal cortex signals decision outcomes to sensory cortex during behavioral adaptations. *Nature Communications* **14**, 3552 (2023).

[20] Liu, D. *et al.* Orbitofrontal control of visual cortex gain promotes visual associative learning. *Nature communications* **11**, 2784 (2020).

[21] Poort, J. *et al.* Learning enhances sensory and multiple non-sensory representations in primary visual cortex. *Neuron* **86**, 1478–1490 (2015).

[22] Palmer, L. M. *et al.* Nmda spikes enhance action potential generation during sensory input. *Nature neuroscience* **17**, 383–390 (2014).

[23] Kuhlman, S. J., O'Connor, D. H., Fox, K. & Svoboda, K. Structural plasticity within the barrel cortex during initial phases of whisker-dependent learning. *Journal of Neuroscience* **34**, 6078–6083 (2014).

[24] Goltstein, P. M., Coffey, E. B., Roelfsema, P. R. & Pennartz, C. M. In vivo two-photon ca2+ imaging reveals selective reward effects on stimulus-specific assemblies in mouse visual cortex. *Journal of Neuroscience* **33**, 11540–11555 (2013).

[25] Guo, L., Ponvert, N. D. & Jaramillo, S. The role of sensory cortex in behavioral flexibility. *Neuroscience* **345**, 3–11 (2017).

[26] Henschke, J. U. *et al.* Reward association enhances stimulus-specific representations in primary visual cortex. *Current Biology* **30**, 1866–1880 (2020).

[27] Nassar, M. R., McGuire, J. T., Ritz, H. & Kable, J. W. Dissociable forms of uncertainty-driven representational change across the human brain. *Journal of Neuroscience* **39**, 1688–1698 (2019).

[28] Biró, B. *et al.* The neural correlates of context driven changes in the emotional response: An fmri study. *Plos one* **17**, e0279823 (2022).

[29] Wang, Y. *et al.* Anatomical, physiological and molecular properties of martinotti cells in the somatosensory cortex of the juvenile rat. *The Journal of physiology* **561**, 65–90 (2004).

[30] Murayama, M. *et al.* Dendritic encoding of sensory stimuli controlled by deep cortical interneurons. *Nature* **457**, 1137–1141 (2009).

[31] Makino, H. & Komiyama, T. Learning enhances the relative impact of top-down processing in the visual cortex. *Nature neuroscience* **18**, 1116–1122 (2015).

[32] Takahashi, N. *et al.* Active dendritic currents gate descending cortical outputs in perception. *Nature Neuroscience* **23**, 1277–1285 (2020).

[33] Rabinovich, R. J., Kato, D. D. & Bruno, R. M. Learning enhances encoding of time and temporal surprise in mouse primary sensory cortex. *Nature Communications* **13**, 5504 (2022).

[34] Rodgers, C. C. & DeWeese, M. R. Neural correlates of task switching in prefrontal cortex and primary auditory cortex in a novel stimulus selection task for rodents. *Neuron* **82**, 1157–1170 (2014).

[35] Nogueira, R. *et al.* Lateral orbitofrontal cortex anticipates choices and integrates prior with current information. *Nature Communications* **8**, 14823 (2017).

[36] Ottenheimer, D. J., Hjort, M. M., Bowen, A. J., Steinmetz, N. A. & Stuber, G. D. A stable, distributed code for cue value in mouse cortex during reward learning. *elife* **12**, RP84604 (2023).

[37] Lee, S., Kruglikov, I., Huang, Z. J., Fishell, G. & Rudy, B. A disinhibitory circuit mediates motor integration in the somatosensory cortex. *Nature neuroscience* **16**, 1662–1670 (2013).

[38] Pi, H.-J. *et al.* Cortical interneurons that specialize in disinhibitory control. *Nature* **503**, 521–524 (2013).

[39] Karnani, M. M. *et al.* Opening holes in the blanket of inhibition: localized lateral disinhibition by vip interneurons. *Journal of neuroscience* **36**, 3471–3480 (2016).

[40] Canto-Bustos, M., Friason, F. K., Bassi, C. & Oswald, A.-M. M. Disinhibitory circuitry gates associative synaptic plasticity in olfactory cortex. *Journal of Neuroscience* **42**, 2942–2950 (2022).

[41] Letzkus, J. J. *et al.* A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* **480**, 331–335 (2011).

[42] van Versendaal, D. & Levelt, C. N. Inhibitory interneurons in visual cortical plasticity. *Cellular and Molecular Life Sciences* **73**, 3677–3691 (2016).

[43] Khan, A. G. *et al.* Distinct learning-induced changes in stimulus selectivity and interactions of gabaergic interneuron classes in visual cortex. *Nature neuroscience* **21**, 851–859 (2018).

[44] Williams, L. E. & Holtmaat, A. Higher-order thalamocortical inputs gate synaptic long-term potentiation via disinhibition. *Neuron* **101**, 91–102 (2019).

[45] Letzkus, J. J., Wolff, S. B. & Lüthi, A. Disinhibition, a circuit mechanism for associative learning and memory. *Neuron* **88**, 264–276 (2015).

[46] Szadai, Z. *et al.* Cortex-wide response mode of vip-expressing inhibitory neurons by reward and punishment. *Elife* **11**, e78815 (2022).

[47] Chen, J. L., Carta, S., Soldado-Magraner, J., Schneider, B. L. & Helmchen, F. Behaviour-dependent recruitment of long-range projection neurons in somatosensory cortex. *Nature* **499**, 336–340 (2013).

[48] Chen, J. L. *et al.* Pathway-specific reorganization of projection neurons in somatosensory cortex during learning. *Nature neuroscience* **18**, 1101–1108 (2015).

[49] Yang, H., Kwon, S. E., Severson, K. S. & O'connor, D. H. Origins of choice-related activity in mouse somatosensory cortex. *Nature neuroscience* **19**, 127–134 (2016).

[50] Burke, K. A., Franz, T. M., Miller, D. N. & Schoenbaum, G. The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. *Nature* **454**, 340–344 (2008).

[51] Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H. & Rushworth, M. F. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron* **65**, 927–939 (2010).

[52] Rolls, E. T. The orbitofrontal cortex and reward. *Cerebral cortex* **10**, 284–294 (2000).

[53] Banerjee, A., Wang, B. A., Teutsch, J., Helmchen, F. & Pleger, B. Analogous cognitive strategies for tactile learning in the rodent and human brain. *Progress in Neurobiology* **222**, 102401 (2023).

[54] Stalnaker, T. A., Cooch, N. K. & Schoenbaum, G. What the orbitofrontal cortex does not do. *Nature neuroscience* **18**, 620–627 (2015).

[55] Knudsen, E. B. & Wallis, J. D. Taking stock of value in the orbitofrontal cortex. *Nature Reviews Neuroscience* **23**, 428–438 (2022).

[56] Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).

[57] Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* **91**, 1402–1412 (2016).

[58] Ma, F., Zhang, L. & Zhou, J. Event-specific and persistent representations for contextual states in orbitofrontal neurons. *Current Biology* (2024).

[59] Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience* **19**, 366–374 (2016).

[60] Masset, P., Ott, T., Lak, A., Hirokawa, J. & Kepecs, A. Behavior-and modality-general representation of confidence in orbitofrontal cortex. *Cell* **182**, 112–126 (2020).

[61] Friston, K., Kilner, J. & Harrison, L. A free energy principle for the brain. *Journal of physiology-Paris* **100**, 70–87 (2006).

[62] Yu, J. A. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).

[63] Grossman, C. D., Bari, B. A. & Cohen, J. Y. Serotonin neurons modulate learning rate through uncertainty. *Current Biology* **32**, 586–599 (2022).

[64] Wilmes, K. A. & Clopath, C. Inhibitory microcircuits for top-down plasticity of sensory representations. *Nature communications* **10**, 5055 (2019).

[65] Krabbe, S. *et al.* Adaptive disinhibitory gating by vip interneurons permits associative learning. *Nature neuroscience* **22**, 1834–1843 (2019).

[66] Lee, S., Hjerling-Leffler, J., Zagha, E., Fishell, G. & Rudy, B. The largest group of superficial neocortical gabaergic interneurons expresses ionotropic serotonin receptors. *Journal of Neuroscience* **30**, 16796–16808 (2010).

[67] Férézou, I. *et al.* 5-ht3 receptors mediate serotonergic fast synaptic excitation of neocortical vasoactive intestinal peptide/cholecystokinin interneurons. *Journal of Neuroscience* **22**, 7389–7397 (2002).

[68] Clarke, H., Walker, S., Dalley, J., Robbins, T. & Roberts, A. Cognitive inflexibility after prefrontal serotonin depletion is behaviorally and neurochemically specific. *Cerebral cortex* **17**, 18–27 (2007).

[69] Barlow, R. L. *et al.* Markers of serotonergic function in the orbitofrontal cortex and dorsal raphé nucleus predict individual variation in spatial-discrimination serial reversal learning. *Neuropsychopharmacology* **40**, 1619–1630 (2015).

[70] Alsiö, J. *et al.* Serotonergic innervations of the orbitofrontal and medial-prefrontal cortices are differentially involved in visual discrimination and reversal learning in rats. *Cerebral Cortex* **31**, 1090–1105 (2021).

[71] Iigaya, K., Fonseca, M. S., Murakami, M., Mainen, Z. F. & Dayan, P. An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. *Nature communications* **9**, 2477 (2018).

[72] Matias, S., Lottem, E., Dugué, G. P. & Mainen, Z. F. Activity patterns of serotonin neurons underlying cognitive flexibility. *Elife* **6**, e20552 (2017).

[73] De Filippo, R. & Schmitz, D. Synthetic surprise as the foundation of the psychedelic experience. *Neuroscience & Biobehavioral Reviews* 105538 (2024).

[74] Chavlis, S. & Poirazi, P. Drawing inspiration from biological dendrites to empower artificial neural networks. *Current opinion in neurobiology* **70**, 1–10 (2021).

[75] Verbeke, P. & Verguts, T. Using top-down modulation to optimally balance shared versus separated task representations. *Neural networks* **146**, 256–271 (2022).

[76] Iyer, A. *et al.* Avoiding catastrophe: Active dendrites enable multi-task learning in dynamic environments. *Frontiers in neurorobotics* **16**, 846219 (2022).

[77] Haider, P. *et al.* Latent equilibrium: A unified learning theory for arbitrarily fast computation with arbitrarily slow neurons. *Advances in neural information processing systems* **34**, 17839–17851 (2021).

[78] Senn, W. *et al.* A neuronal least-action principle for real-time learning in cortical circuits. *eLife* **12** (2024).

[79] Max, K. *et al.* Learning efficient backprojections across cortical hierarchies in real time. *Nature Machine Intelligence* 1–12 (2024).

[80] Ellenberger, B. *et al.* Backpropagation through space, time, and the brain. *arXiv preprint arXiv:2403.16933* (2024).

[81] Banerjee, A., Rikhye, R. V. & Marblestone, A. Reinforcement-guided learning in frontal neocortex: emerging computational concepts. *Current Opinion in Behavioral Sciences* **38**, 133–140 (2021).

[82] Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience* **30**, 12366–12378 (2010).

[83] Soltani, A. & Izquierdo, A. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience* **20**, 635–644 (2019).

[84] Pachitariu, M. *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *BioRxiv* 061507 (2016).

Acknowledgements

This works was supported by the Swiss National Science Foundation (project grants 170269 and 192617 to FH, Sinergia grant 180316 to FH and WS), the European Union's Horizon 2020 research and innovation programme through the ICEI project under the grant agreement No. 800858 (Human Brain Project, to WS), the European Research Council (ERC Advanced Grant BRAINCOMPATH, project 670757, to FH), a Wellcome Trust institutional strategic award (to AB), a Royal Society research grant (RGS\R2\202155, to AB), and a Wellcome Trust Career Development Award (to AB). AB is affiliated to the Digital Environment Research Institute at QMUL. The authors also wish to thank Johanni Brea and Alireza Modirshanechi for helpful discussions.

Author contributions

MT, AB and WS conceptualized and wrote the paper. Experimental investigations were undertaken by JT and supervised by AB. Computational and mathematical investigations were carried out by MT and supervised by WW and WS. Visualization was performed by MT, JT, AB and WS. Funding was acquired by WW, FH, AB and WS. All authors contributed to reviewing and editing.

Competing interests

The authors declare no competing interests.

Supplementary Information

Contents

Supplementary Methods

1 Inferring action-values based on the outcome prediction network

The action-value estimate $\hat{Q}$ was computed by using an outcome prediction network. For each outcome $\omega \in \{\text{reward}, \text{punishment}\}$, we defined the indicator function Ω_ω, which indicated whether an outcome ω took place ($\Omega_\omega = 1$) or not ($\Omega_\omega = 0$). The estimated probability to encounter outcome ω upon choosing action $a \in \{\text{go}, \text{no}-\text{go}\}$ in response to stimulus representation $\boldsymbol{z}$ is denoted by $\hat{p}_\omega(a, \boldsymbol{z})$. We modeled these estimated outcome probabilities by means of 4 perceptrons, one for each of the 4 action-outcome pairs (a, ω), with a standard logistic activation function φ,

$$\hat{p}_\omega(a, \boldsymbol{z}) = \varphi\Big((\boldsymbol{w}^Q_{a\omega})^{\mathrm{T}} \boldsymbol{z}\Big) \,. \tag{14}$$

These probability estimators were trained online, each time an action a' was taken according to

$$\Delta w^Q_{a'\omega j} = \eta^Q \varphi'\Big((\boldsymbol{w}^Q_{a'\omega})^{\mathrm{T}} \boldsymbol{z}\Big) \Big(\Omega_\omega - \hat{p}_\omega(a', \boldsymbol{z})\Big) \, z_j \,, \tag{15}$$

where φ' denotes the derivative of φ. Because the state-action-outcome triples $(\boldsymbol{z}, a, \omega)$ define the rule, the outcome predictions $\hat{p}_\omega(a, \boldsymbol{z})$ represent the internal understanding about the rule (and more generally about the context). With these outcome predictors, $\hat{Q}$ can be expressed by a sum weighted according to the value $R_\omega \in \{-1, 1\}$ of each outcome

$$\hat{Q}(a, z) = \sum_\omega R_\omega \, \hat{p}_\omega(a, \boldsymbol{z}) = \hat{p}_{\text{reward}}(a, \boldsymbol{z}) - \hat{p}_{\text{punish}}(a, \boldsymbol{z}) \,. \tag{16}$$

For our two outcomes, we set $R_{\text{reward}} = 1$ and $R_{\text{punish}} = -1$, but the model can also be generalized to more outcome types.

2 Action selection by policy gradient and action-value estimate

Action selection was simulated to be based on the sensory representation $\boldsymbol{z}$. The policy network consisted of a perceptron with sigmoidal transfer function ϕ, which produced the action selection probability π^{PG},

$$\pi^{\text{PG}}(a_{\text{go}} \mid \boldsymbol{z}) = \phi((\boldsymbol{w}^{\text{PG}})^{\mathrm{T}} \boldsymbol{z}) \tag{17}$$

$$\pi^{\text{PG}}(a_{\text{nogo}} \mid \boldsymbol{z}) = 1 - \pi^{\text{PG}}(a_{\text{go}} \mid \boldsymbol{z}) \tag{18}$$

$$\phi = \epsilon_\phi + \frac{(1 - 2\epsilon_\phi)}{1 + b_\phi \exp(-x)} \,, \tag{19}$$

where ϵ_ϕ and b_ϕ are constants. The network learned through policy gradient [85] by minimizing the loss

$$L^{\text{PG}} = -R \log(\pi(a'|\boldsymbol{z})) \,, \tag{20}$$

where a' is the chosen action and R is the reward amplitude of the outcome observed. This loss was minimized by online gradient descent with learning rate η^{PG} according to the learning rule

$$\Delta \boldsymbol{w}^{\mathrm{PG}} = \eta^{\mathrm{PG}} \frac{dL^{\mathrm{PG}}}{d\boldsymbol{w}^{\mathrm{PG}}} \,. \tag{21}$$

Since a pure policy gradient network approach can lead to an excessively exploitative action selection with very little exploration after convergence, it can be inflexible when facing non-stationary conditions such as rule reversals (Extended Data Fig. S4e). We therefore combined it with a more exploratory policy π^Q, which is computed by weighing the predicted outcomes $\hat{Q}(a, \boldsymbol{z})$, with a softmax transfer function parameterized by a temperature constant T^{Q},

$$\pi^{\mathrm{Q}}(a|\boldsymbol{z}) = \frac{\exp\left(\hat{Q}(a, \boldsymbol{z})/T^{\mathrm{Q}}\right)}{\sum_i \exp\left(\hat{Q}(a_i, \boldsymbol{z})/T^{\mathrm{Q}}\right)} \,, \tag{22}$$

where $\hat{Q}(a, \boldsymbol{z})$ is given by equation (16) and the sum extends across the two actions $a_i \in \{a_{\mathrm{GO}}, a_{\mathrm{NG}}\}$, corresponding to lick or no lick, respectively go or no-go. Both the exploratory policy π^Q and the exploitative policy π^{PG} were incorporated into the final policy π, which consisted of the confidence weighted convex combination (Extended Data Fig. S4a),

$$\pi(a|\boldsymbol{z}) = c \cdot \pi^{\mathrm{PG}}(a|\boldsymbol{z}) + (1 - c) \cdot \pi^{Q}(a|\boldsymbol{z}) \,, \tag{23}$$

where c is the prediction confidence, see equation (7). The overall policy π was shown to learn effectively upon reversal learning and therefore ensured that any lack of adaptability would not be caused by the inflexibility in the action selection network (Extended Data Fig. S4e). Depending on the purpose of the modeling, we used different algorithms for the action selection. In most cases, the convex policy was used as described in equation (11). For Fig. 1 however, the purpose was to highlight the impact of the sensory representation z and gain modulation for a standard policy, which is why the pure policy network π^{PG} was used (for completeness purposes Fig. 1 is replicated with the convex policy π in Extended Data Fig. S4). For the first reversal simulations — which illustrated the dynamics of uncertainty, confidence and context-prediction error (Fig. 2b,c,f) — we imposed a performance learning curve, dictated by a fixed evolution of the correct action probability $p(a_{\mathrm{opt}})$ (Extended Data Fig. S1d).

3 Model combinations

In this manuscript, we implemented different combinations of action-selection policies and gain modulation mechanisms (Extended Data Fig. S1). The action-selection policy was one of: policy gradient, fixed performance trace (fixed policy), convex policy. The gain modulation was one of: none, only reward-guided amplification without lOFC, amplification with lOFC targeting SST interneurons in S1, amplification with lOFC switching the context representation. Extended Data Table 1 summarizes, which model was shown in which figure. While multiple model combinations were implemented for completeness purposes, three were central to our main findings. 1) To demonstrate the advantage provided by top-down excitation in stationary environments, we combined classical policy gradient for action selection with apical gain amplification of sensory neurons based on reward association (Extended Data Fig. S1a). 2) To illustrate the second-order statistics of reward-prediction errors in a non-stationary environment (without using the full model adapted for the purpose of reversal learning), we used a fixed policy performance trace (Extended Data Fig. S1b). This allowed to generate a realistic evolution of the reward sampling distribution and study the dynamics of confidence loss and context-prediction error during reversal learning (Extended Data Fig. S1d). 3) Finally, the full model with lOFC feedback to S1 SST interneurons and in some instances affecting the context representation used a confidence-weighted convex policy (Extended Data Fig. S1c). The convex policy ensured that any lack of reversal performance would not be due to deficits in the flexibility of the policy (Extended Data Fig. S4), but could be attributed to failing to re-adapt the sensory representation.

4 Parameter values

The parameters $\eta^V, \eta^\pi, \eta^Q, r_0^{\mathrm{VIP}}$ and w^{SST} were optimized by grid search, while all remaining parameters had fixed values, which are reported in Extended Data Table 2. The parameter optimization took place by testing different sets of parameters values taken from logarithmic grid. Each set of parameters was tested with 64 pseudo-random initializations. After the best set of parameters was determined, a new tighter logarithmic grid of parameter values was created surrounding the best parameter configurations. Each new set of parameters was then tested to determine, if they could improve on the best configuration. This process was repeated a second time with yet a tighter grid and the final best parameter configurations was chosen, see Extended Data Algorithm 2. The criterion for the best set of parameters was usually the number of trials to reach expert performance after one reversal. In some instances, like in Fig. 1, we optimized for the final performance for the original rule (see Learning perf. in Extended Data Table 3) and in one example, we optimized for the number of trials to reach expert performance in the original rule (Extended

Data Fig. S6).
The optimized values are reported in Extended Data Table 3. While many different parameter optimizations were performed, only two are relevant for the main results. 1) For the results in Fig. 1, we optimized the parameters with respect to the final learning performance (without any rule reversal). 2) For the remaining main Figures, we first optimized the model without lOFC to achieve expert performance after one reversal in the least number of trials possible. This allowed as to fix the policy related parameters η^{π} and η^{Q}, before optimizing the remaining parameters related to the top-down modulation of the sensory representation. While this 'mixed' optimization strategy did not target the theoretically best parameter configuration for the full model with lOFC (as illustrated in Extended Data Fig. S6), it prevented any potential bias towards the model with lOFC, when comparing the model with and without lOFC (Fig 3).

Supplementary Equations

1 Gains modulate neuron-specific learning rates

For a rate-based neuron j, the multiplicative gain g_j modulates the pre-gain activation x_j to produce the firing rate

$$z_j = g_j \lfloor x_j \rfloor \ . \tag{24}$$

At the same time, most synaptic weight learning rules — from Hebb to error backpropagation — are almost universally proportional to the presynaptic activation. As a consequence, increasing the multiplicative gain of a neuron, raises the rate of change of its efferent synapses. From the perspective of the weight learning rule, the learning rate of a synapse w_{ij} can be viewed as being modulated by the multiplicative gain g_j of the afferent neuron j,

$$\Delta w_{ij} \propto \eta_0 \, z_j = \eta_0 \, g_j \lfloor x_j \rfloor = \eta_j \lfloor x_j \rfloor \ . \tag{25}$$

Here, η_0 corresponds to the traditional global learning rate and $\eta_j = \eta_0 \, g_j$ is the learning rate of synapses activated by a neuron j. Therefore, by increasing the gain of task-relevant compared to task-irrelevant sensory neurons, it is possible to help the synapses of the policy network to learn more strongly from the task-relevant information. On the other side, reducing the gain will reduce the plasticity. Hence SST interneurons activated by lOFC upon reversal can protect policy weights from being unlearned.

2 Apical amplification with respect to task-relevance

In this section, we will elaborate on our model for learning sensory representations that exhibit heightened responses to task-relevant information. For this, we assume that in biological neural networks, the main mechanism to produce selective sensory representations is not based on the synaptic plasticity of feed-forward pathways. Rather, in order to keep a wide range of features available for multiple different contexts, the brain needs to increase the gain of task-relevant neurons via top-down afferents. We further assume that this gain modulation occurs by targeting the apical dendrites of sensory pyramidal neurons. With this in mind, we can set the following objective for our top-down gain modulation mechanism:

→ **High task-relevance of a stimulus should lead to a high gain and low task-relevance a low gain.**

This objective can be formalized and implemented in different ways. We opted to define the task-relevance of a stimulus i in terms of its contribution to the outcome prediction $\hat{V}$,

$$\hat{V}(\boldsymbol{z}) \approx \sum_i \frac{d\hat{V}}{dz_i} \cdot z_i = \sum_i \Delta \hat{V}_i \ . \tag{26}$$

Here the approximation becomes an equality in the case of a linear dependence, which is the case in our model. The term $\Delta \hat{V}_i$ can be interpreted as the reward prediction value gained by the model from the activity of neuron i. For apical excitation, we are interested in the the positive part $v_i = \lfloor \Delta \hat{V}_i \rfloor$, of which the expectation $\mathbb{E}[v_i]$ makes our definition of task-relevance. In order to produce an apical gain amplification based on the task-relevance of a neuron i, we applied the plasticity rule

$$\Delta w_{ij}^{\mathrm{ap}} = \eta^{\mathrm{ap}} \left(v_i - r^{\mathrm{SST}} \right) x_j^c \ . \tag{27}$$

By analyzing this update rule, we can derive the convergence state of the apical excitation for a neuron i. For this we will consider the setting, where $j \in \{1\}$ and $x_j^c = 1$. In expectation the synaptic plasticity can hence be formulated as

$$\mathbb{E}\left[\Delta w_{ij}^{\mathrm{ap}}\right] = \mathbb{E}\left[\eta^{\mathrm{ap}} \left(v_i - r^{\mathrm{SST}} \right)\right] \propto \mathbb{E}\left[v_i\right] - \mathbb{E}\left[r^{\mathrm{SST}}\right] \ . \tag{28}$$

We find that the plasticity rule lets the apical excitation transmitted by the apical synapses w_{ij}^{ap} grow large for neurons with a task-relevance greater than the average SST activity, and shrink small otherwise,

$$\mathbb{E}\left[w_{ij}^{\text{ap}}\right] = \begin{cases} w_{\max}^{\text{ap}}; & \text{if } \mathbb{E}\left[v_i\right] > \mathbb{E}\left[r^{\text{SST}}\right] \\ w_{\min}^{\text{ap}}; & \text{if } \mathbb{E}\left[v_i\right] < \mathbb{E}\left[r^{\text{SST}}\right] \end{cases} . \tag{29}$$

This confirms that $\mathbb{E}[\lfloor \Delta \hat{V}_i \rfloor]$ can be considered as out effective definition of task-relevance. The rectifier function $\lfloor \cdot \rfloor$ is used to prevent the explicit apical inhibition of neurons that are predictive of negative outcomes. Instead, these negative contributions will have the same effect on the gain modulation those of neurons, which are irrelevant for the outcome prediction (i.e. neurons with $\Delta \hat{V}_i = 0$). Finally, one can note that the SST activity not only reduces immediate apical activity, but also drives synaptic weight depression in the apical dendrites [86] and thereby raises the task-relevance threshold to drive gain amplification. The SST interneuron activity r^{SST} is excited by the lOFC, which increases r^{SST} in case of context-prediction error, see equation (10). In consequence, following a rule reversal, the lOFC raises the task-relevance threshold for the consolidation of top-down apical amplification.

3 Distinguishing sources of uncertainty

It has been demonstrated that the expected squared error of an estimator can be decomposed into three components: an irreducible, a bias and a variance component [87]. Here, we reproduce this result and adapt it to our case, where distracting information and changes of context come into play. We first define the estimation target $Y = f(X) + \epsilon$ with $\mathbb{E}[\epsilon] = 0$ and $Var(\epsilon) = \sigma_\epsilon^2$. These targets are predicted by a stochastic estimator $\hat{f}(X)$ based on the observables X. The expected squared prediction error between the target and the estimator can be interpreted as the predictive uncertainty. Next we will demonstrate how the predictive uncertainty can be decomposed into three separate sources of uncertainty: (1) an intrinsic and irreducible error (sometimes referred to as *risk* [88, 89]) determined by ϵ, (2) a bias, which is reduced if the expected value of $\hat{f}(X)$ closely matches $f(X)$ and (3) a variance term which arises from the intrinsic stochasticity of $\hat{f}(X)$:

$$\begin{aligned} \mathbb{E}[(Y - \hat{f}(X))^2 | X = x_0] =& \mathbb{E}[(\epsilon + f(x_0) - \hat{f}(x_0))^2] \\ =& \mathbb{E}[\epsilon^2] + \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] \\ =& \sigma_\epsilon^2 + \mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)] + \mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] \\ =& \sigma_\epsilon^2 + \mathbb{E}[(\mathbb{E}[\hat{f}(x_0)] - f(x_0))^2] + \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] \\ & + 2\mathbb{E}[(\mathbb{E}[\hat{f}(x_0)] - f(x_0))(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])] \\ =& \sigma_\epsilon^2 + \left(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\right)^2 + \mathbb{E}[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2] \\ =& \text{Irreducible error} + \text{Bias}^2 + \text{Variance} . \end{aligned} \tag{30}$$

We can also apply this framework to our model, where the estimation target Y represents the reward R of the deterministic go/no-go task and the estimator $\hat{f}$ represents the reward prediction $\hat{Q}$. In this case, the observables X encodes the action as well as the sensory information. Next we will adapt this decomposition to consider distracting inputs and non-stationary changes of context.

Distraction uncertainty. In our model, the reward prediction was deterministic and had no intrinsic stochasticity, meaning the variance term in equation (30) was zero. The reward outcomes were also deterministic, meaning the irreducible error $\sigma_\epsilon^2 = 0$. However the predictor also received task-irrelevant information D. This contribution can be emphasized by splitting off this distracting information D from the observables X as distinct inputs to the estimator $\hat{f}$. Using the same process as in equation (30), we find

$$\begin{aligned} \mathbb{E}[(Y - \hat{f}(X, D))^2 | X = x_0] =& \sigma_\epsilon^2 + (\mathbb{E}[\hat{f}(x_0, D)] - f(x_0))^2 + \mathbb{E}[(\hat{f}(x_0, D) - \mathbb{E}[\hat{f}(x_0, D)])^2] \\ =& \text{Irreducible error} + \text{Bias}^2 + \text{Distraction} . \end{aligned} \tag{31}$$

Under this formulation, the top-down amplification of task-relevant signals can be understood as a means to reduce the distraction component of the prediction uncertainty.

Contextual uncertainty. So far we assumed a single stationary context. In order to consider the possibility of a change of context, we just need to introduce the dependence of the target Y on a current context c^*, while the model assumes the context c^m,

$$Y = f(X, c^*) + \epsilon . \tag{32}$$

We can now decompose the predictive uncertainty in a non-stationary environment. In case the correct context is assumed ($c^m = c^*$),

$$\begin{aligned}\mathbb{E}[(Y - \hat{f}(X, c^*))^2 | X = x_0] =& \sigma_\epsilon^2 + (\mathbb{E}[\hat{f}(x_0, c^*)] - f(x_0, c^*))^2 + \mathbb{E}[\big(\hat{f}(x_0, c^*) - \mathbb{E}[\hat{f}(x_0, c^*)]\big)^2] \\ =& \text{Irreducible error} + \text{Bias}^2 + \text{Variance} \, .\end{aligned} \tag{33}$$

This is similar to what we found in the stationary case, see equation (30). However, if the wrong context is assumed ($c^m = c^\dagger \neq c^*$), we find that the bias term will be affected by the contextual mismatch

$$\begin{aligned}\mathbb{E}[(Y - \hat{f}(x_0, c^\dagger))^2 | X = x_0] =& \sigma_\epsilon^2 + (\mathbb{E}[\hat{f}(x_0, c^\dagger)] - f(x_0, c^*))^2 + \mathbb{E}[\big(\hat{f}(x_0, c^\dagger) - \mathbb{E}[\hat{f}(x_0, c^\dagger)]\big)^2] \\ =& \text{Irreducible error} + \text{Context bias}^2 + \text{Variance} \, .\end{aligned} \tag{34}$$

Since the incorrect context assumption would create a strong bias, we refer to this term as context bias to distinguish it from the bias in the stationary case.

Unexpected uncertainty. Next we will analyse the dynamics of the prediction uncertainty by looking at the mismatch between the true uncertainty σ^2 and the expected uncertainty $\hat{\sigma}^2$ before and after a change of context. So far, to simplify the notation, we only demonstrated how to decompose the conditional expected squared prediction error given the observables x_0. However the analogous decomposition can be made for the full expectation (taken over the intrinsic stochasticity of $\hat{f}$ and over ϵ, but also over the observables X)

$$\sigma^2(c^m) = \mathbb{E}[(Y - \hat{f}(X, c^m))^2] \, , \tag{35}$$

which is the estimation target of the expected prediction uncertainty $\hat{\sigma}^2$, see equation (6). Before a change of context — in case the correct context is assumed ($c^m = c^*$) and the uncertainty estimator $\hat{\sigma}^2$ has converged, $\hat{\sigma}^2 = \sigma^2(c^m)$ — we get

$$\begin{aligned}\hat{\sigma}^2 = \sigma^2(c^*) =& \mathbb{E}[(Y - \hat{f}(X, c^*))^2] \\ =& \int \mathbb{E}[(Y - \hat{f}(x_0, c^*))^2 | X = x_0] P(X = x_0) dx_0 \\ =& \int (\text{Irreducible error} + \text{Bias}^2 + \text{Variance}) P(X = x_0) dx_0 \, .\end{aligned} \tag{36}$$

In a scenario, in which learning of $\hat{f}$ has converged successfully ($\text{Bias}^2 = 0$), a sudden change of context would increase the squared error by inducing a context bias, see equation (34). This increase in uncertainty can be quantified by the unexpected uncertainty, which is obtained by subtracting the true uncertainty just after a change of context $\sigma^2(c^\dagger)$, with the expected uncertainty just before the change of context $\hat{\sigma}^2 = \sigma^2(c^*)$, equation (36). Indeed, we find that the unexpected increase in uncertainty, which can be attributed to the change of context, equates to

$$\begin{aligned}\sigma^2(c^\dagger) - \hat{\sigma}^2 =& \mathbb{E}[(Y - \hat{f}(x_0, c^\dagger))^2] - \hat{\sigma}^2 \\ =& \int \mathbb{E}[(Y - \hat{f}(x_0, c^\dagger))^2 | X = x_0] P(X = x_0) dx_0 - \sigma^2(c^*) \\ =& \int (\text{Irreducible error} + \text{Context bias}^2 + \text{Variance}) P(X = x_0) dx_0 \\ & - \int (\text{Irreducible error} + \text{Bias}^2 + \text{Variance}) P(X = x_0) dx_0 \\ =& \int (\text{Context bias}^2 - \text{Bias}^2) P(X = x_0) dx_0 \, .\end{aligned} \tag{37}$$

In equation (37), we assumed that the variance term was independent on the assumed context C, which is trivial in case the estimator $\hat{f}$ is not stochastic at all. Finally, if learning of $\hat{f}$ had perfectly converged ($\text{Bias}^2 = 0$) prior to the change of context,

$$\sigma^2(c^\dagger) - \hat{\sigma}^2 = \sigma^2(c^\dagger) - \sigma^2(c^*) = \int \text{Context bias}^2 P(X = x_0) dx_0 \, . \tag{38}$$

Therefore, in an idealized setting, the unexpected increase in uncertainty following the change of context becomes equal to the contextual uncertainty. Hence, unexpected uncertainty can be used to estimate contextual uncertainty.

Finally, we can note that while we didn't explicitly write the effect of the distracting inputs for this non-stationary formulation, their contribution is also part of the Context bias^2 term.

4 Explained and unexplained variance

Predictive models are commonly evaluated in terms of the fraction of variance unexplained or variation explained. Here we reiterate this formalism to provide perspective on the estimates of predictive uncertainty $\hat{\sigma}^2$ and confidence c, see equation (7). The expected squared prediction error σ^2 between a prediction target R and a model prediction Q is a general metric of the prediction loss,

$$\sigma^2 = \mathbb{E}[(R - \hat{Q})^2] \,. \tag{39}$$

It is commonly also referred to as the residual variance or unexplained variance. This can be compared to the intrinsic variance of the target

$$\mathrm{Var}(R) = \mathbb{E}[(R - \mathbb{E}[R])^2] \,. \tag{40}$$

By taking the difference between the intrinsic variance and the unexplained variance, we get the variance explained VE (also referred to as explained variation)

$$\mathrm{VE} = \lfloor \mathrm{Var}(R) - \sigma^2 \rfloor \,. \tag{41}$$

The VE quantifies how much the prediction error is reduced by the model $\hat{Q}$ compared to using the unconditional expected value $\mathbb{E}[R]$ as an estimate. By dividing the unexplained variance σ^2 with the intrinsic variance, we get the fraction of variance unexplained FVU,

$$\mathrm{FVU} = \sigma^2 / \mathrm{Var}(R) \,. \tag{42}$$

The analogous can be done to get the fraction of variance explained FVE, commonly referred to as coefficient of determination $\mathcal{R}^2$,

$$\mathcal{R}^2 = \mathrm{VE}/\mathrm{Var}(R) = \left\lfloor 1 - \frac{\sigma^2}{\mathrm{Var}(R)} \right\rfloor = \lfloor 1 - \mathrm{FVU} \rfloor = \mathrm{FVE} \,. \tag{43}$$

With this formalism at hand, the expected uncertainty $\hat{\sigma}^2$, which is an online estimator of σ^2, can be viewed as an internal representation of the unexplained variance. On the other hand, the prediction confidence

$$c = \left\lfloor 1 - \frac{\hat{\sigma}^2}{\sigma^2_{\max}} \right\rfloor^2 \,, \tag{44}$$

as defined in equation (7), represents an internal estimate of FVE^2, the square of the fraction of variance explained, see equation (43). The rectifier function ensures that the confidence remains zero, even if the model $\hat{Q}$ is a worse estimator than $\mathbb{E}[R]$. While this situation would typically not be considered in stationary contexts, changes of context can lead the squared prediction error σ^2 to rise above the intrinsic variance.

Note: Alternatively, we could have first introduced an objective prediction confidence as $c = \mathrm{FVE}^2$ and then the subjective estimated confidence as $\hat{c} = \left\lfloor 1 - \hat{\sigma}^2/\sigma^2_{\max} \right\rfloor^2$. However, for simplicity we directly defined the internal representation of the prediction confidence as c, hence dropping the $\hat{}$ from $\hat{c}$, see equation (44).

5 Computing the reward variance

In the go/no-go sensory discrimination task, the distribution of the rewards R depend on the action selection probabilities. While it would be possible to continuously estimate the variance of the rewards for a changing distribution of action selection probabilities, this would require an online estimator. For simplicity, we use a fixed value $\sigma^2_{\max}$, which was set to the variance in case of a uniformly random action selection policy $\pi(a_{\mathrm{GO}}|s_1) = \pi(a_{\mathrm{GO}}|s_2) = \pi(a_{\mathrm{NG}}|s_1) = \pi(a_{\mathrm{NG}}|s_2) = 0.5$. The value $\sigma^2_{\max} = 0.5$ can be shown by

$$\begin{aligned}
\sigma^2_{\max} &= \mathbb{E}[(R - \mathbb{E}[R])^2] = \mathbb{E}[R^2] && (45.1)\\
&= \sum_R p(R) R^2 = \sum_{s \in \{s_1, s_2\}} p(s) \sum_{a \in \{a_{\mathrm{GO}}, a_{\mathrm{NG}}\}} \pi(a|s) \sum_R p(R|s,a) R^2 && (45.2)\\
&= \sum_{s \in \{s_1, s_2\}} \frac{1}{2} \sum_{a \in \{a_{\mathrm{GO}}, a_{\mathrm{NG}}\}} \frac{1}{2} \sum_R p(R|s,a) R^2 && (45.3)\\
&= \frac{1}{4} \sum_s \sum_a \sum_R p(R|s,a) R^2 && (45.4)
\end{aligned}$$

$$= \frac{1}{4}(R(\text{Hit})^2 + R(\text{CR})^2 + R(\text{FA})^2 + R(\text{Miss})^2) \tag{45.5}$$

$$= \frac{1}{4}(1^2 + 0^2 + (-1)^2 + 0^2) = 0.5\ . \tag{45.6}$$

Equation (45.3) used the uniformity of stimulus sampling, meaning $p(s_1) = p(s_2) = 0.5$. Equations (45.5) and (45.6) used that, conditioned on action and stimulus, the reward outcomes are deterministic ($p(R|s,a) \in \{0,1\}$), and the rewards R were $\{1,-1,0\}$ for Hit, FA and either CR or Miss respectively. Equation (45.1) used $\mathbb{E}[R] = 0$. This can be demonstrated by

$$\begin{aligned}
\mathbb{E}[R] &= \sum_R p(R)R \\
&= \sum_s p(s) \sum_a \pi(a|s) \sum_R p(R|s,a)R \\
&= \frac{1}{4} \sum_s \sum_a \sum_R p(R|s,a)R \\
&= \frac{1}{4}(1 + 0 - 1 + 0) = 0\ .
\end{aligned} \tag{46}$$

Note: in the same way that $\hat{\sigma}^2$ can be simultaneously interpreted both as the estimate of the variance unexplained by the predictor and as the expected prediction uncertainty, $\sigma^2_{\max}$ can be viewed under both lenses. On one side, it represents the variance of the reward, which can be used to compute the fraction of variance unexplained that appears in the coefficient of determination. On the other side, $\sigma^2_{\max}$ can be viewed as the prediction uncertainty of the most simple estimator $\hat{R} = \mathbb{E}[R]$. This explains, why the predictive confidence is zero, if the expected uncertainty $\hat{\sigma}^2$ is greater than $\sigma^2_{\max}$, see equation (7).

6 Derivation of the confidence loss

The OFC activity at stimulus time is simulated as a prediction of the confidence loss CL. Confidence loss is modeled as being elicited by unexpected outcomes leading to a decrease in predictive confidence. The change in predictive confidence c is quantified by taking the difference in confidence between trial $t+1$ and trial t:

$$\Delta c = c_{t+1} - c_t = \left\lfloor 1 - \frac{\hat{\sigma}^2_{t+1}}{\sigma^2_{\max}} \right\rfloor^2 - \left\lfloor 1 - \frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}} \right\rfloor^2 . \tag{47}$$

We ask how a change $\Delta\hat{\sigma}^2 = \hat{\sigma}^2_{t+1} - \hat{\sigma}^2_t$ in the predictive uncertainty maps to a change Δc in the predictive confidence. For this, we calculated the derivative of c with respect to $\hat{\sigma}^2$,

$$\frac{dc}{d\hat{\sigma}^2} = \frac{d}{d\hat{\sigma}^2} \left\lfloor 1 - \frac{\hat{\sigma}^2}{\sigma^2_{\max}} \right\rfloor^2 = - \left\lfloor 1 - \frac{\hat{\sigma}^2}{\sigma^2_{\max}} \right\rfloor \frac{2}{\sigma^2_{\max}} = -\frac{2\sqrt{c}}{\sigma^2_{\max}} . \tag{48}$$

For small $\Delta\hat{\sigma}^2$ we can there for approximate

$$\Delta c \approx \frac{dc}{d\hat{\sigma}^2} \Delta\hat{\sigma}^2 = -\frac{2\sqrt{c}}{\sigma^2_{\max}} \Delta\hat{\sigma}^2 . \tag{49}$$

The same result can be obtained with finite time differences. For notational simplicity we assume that $\hat{\sigma}^2 \in [0, \sigma^2_{\max}]$, so that we can neglect the rectifier function $\lfloor \cdot \rfloor$. We then find

$$\begin{aligned}
\Delta c &= \left(1 - \frac{\hat{\sigma}^2_{t+1}}{\sigma^2_{\max}}\right)^2 - \left(1 - \frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right)^2 \\
&= -2\frac{\hat{\sigma}^2_{t+1} - \hat{\sigma}^2_{t}}{\sigma^2_{\max}} + \left(\frac{\hat{\sigma}^2_{t+1}}{\sigma^2_{\max}}\right)^2 - \left(\frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right)^2 \\
&= -2\frac{\Delta\hat{\sigma}^2}{\sigma^2_{\max}} + \left(\frac{\Delta\hat{\sigma}^2 + \hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right)^2 - \left(\frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right)^2 \\
&= \frac{1}{\sigma^2_{\max}}\left(-2\Delta\hat{\sigma}^2 + \frac{(\Delta\hat{\sigma}^2)^2 + 2\Delta\hat{\sigma}^2\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right) \\
&= \frac{2\Delta\hat{\sigma}^2}{\sigma^2_{\max}}\left(-1 + \frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right) + \frac{(\Delta\hat{\sigma}^2)^2}{\sigma^4_{\max}} \\
&= \frac{2\Delta\hat{\sigma}^2}{\sigma^2_{\max}}\left(-1 + \frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right) + O\left(\left(\Delta\hat{\sigma}^2\right)^2\right) \\
&= -\frac{2\Delta\hat{\sigma}^2}{\sigma^2_{\max}}\left(1 - \frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right) + O\left(\left(\Delta\hat{\sigma}^2\right)^2\right) \ .
\end{aligned} \tag{50}$$

Confidence loss can therefore be approximated as

$$\begin{aligned}
\mathrm{CL} = \lfloor -\Delta c \rfloor &= \left\lfloor \frac{2\Delta\hat{\sigma}^2}{\sigma^2_{\max}}\left(1 - \frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right) - O\left(\left(\Delta\hat{\sigma}^2\right)^2\right)\right\rfloor \\
&\approx \left\lfloor \frac{2\Delta\hat{\sigma}^2}{\sigma^2_{\max}}\left(1 - \frac{\hat{\sigma}^2_{t}}{\sigma^2_{\max}}\right)\right\rfloor = \frac{2}{\sigma^2_{\max}} \cdot \sqrt{c_t} \cdot \lfloor \Delta\hat{\sigma}^2 \rfloor \ .
\end{aligned} \tag{51}$$

In the last step, we again made use of the fact that $\hat{\sigma}^2 \in [0, \sigma^2_{\max}]$ (or respectively that $c_t \in [0, 1]$).

Supplementary References

[85] Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* **8**, 229–256 (1992).

[86] Hayama, T. *et al.* Gaba promotes the competitive selection of dendritic spines by controlling local ca2+ signaling. *Nature neuroscience* **16**, 1409–1416 (2013).

[87] Geman, S., Bienenstock, E. & Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation* **4**, 1–58 (1992).

[88] Payzan-LeNestour, E. & Bossaerts, P. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS computational biology* **7**, e1001048 (2011).

[89] Preuschoff, K., Quartz, S. R. & Bossaerts, P. Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience* **28**, 2745–2752 (2008).

Extended data

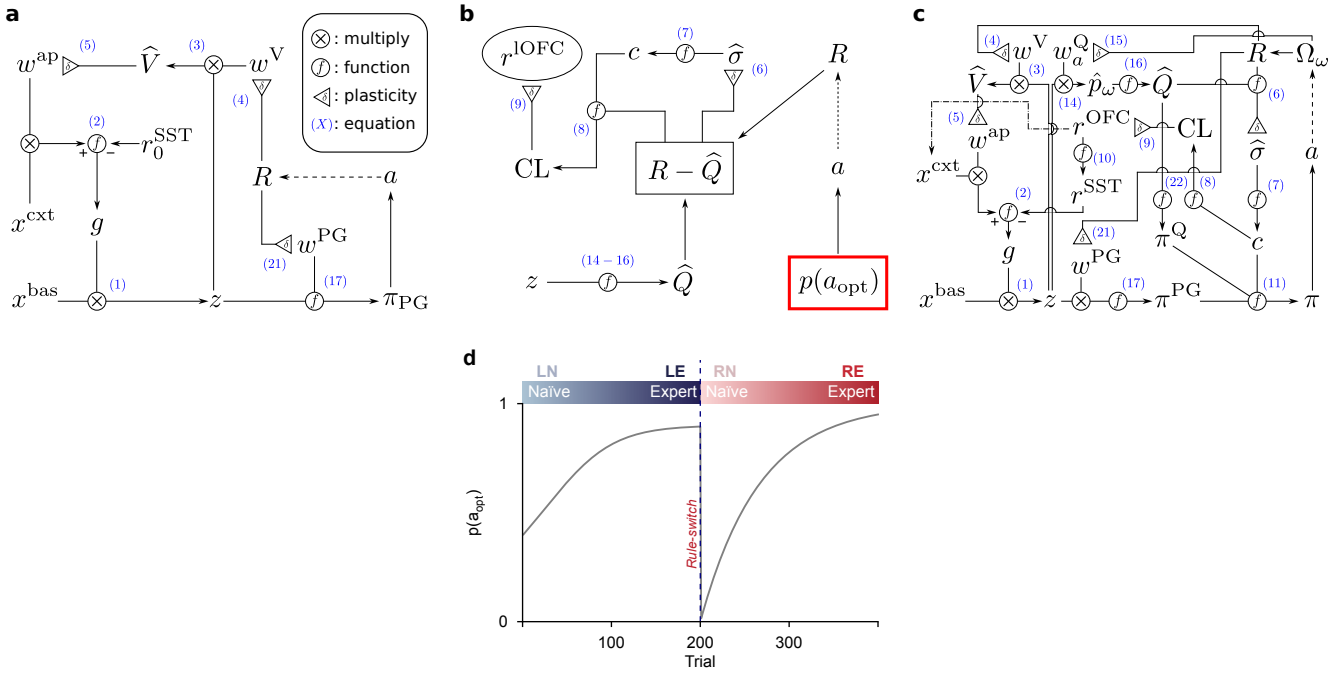

Fig. S1 Network map of the model and its equations. a, Map of the gain modulation apparatus without the lOFC and using pure policy gradient for action selection. This model was used in the simulations for Fig. 1 to demonstrate, how apical gain amplification based on value facilitates online policy gradient learning from sensory representations encoding distracting information. The numbers in blue refer to the equation numbers for mathematical descriptions. **b**, Network describing how the lOFC activity was modeled for the simulations in Fig. 2, at which point the top-down influence of the lOFC to the sensory representation z was not yet included. In order to nonetheless use a realistic evolution of the action selection performance, the policy was replaced by a fixed trace dictating the evolution of the correct action probability $p(a_{\text{opt}})$. **c**, Network schematic of the full model containing the two suggested top-down influences of the lOFC onto the sensory representation (one via SST interneurons and one switching the context representation $\boldsymbol{x}^{\text{cxt}}$). **d**, Fixed evolution of the correct action probability used in the simulations described in panel b.

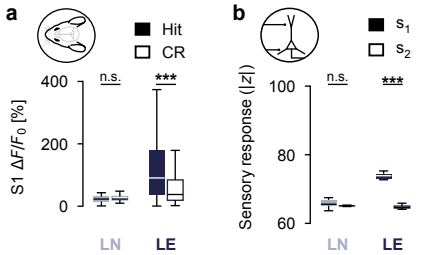

Fig. S2 Reward association amplifies stimulus response. a, Calcium imaging of mouse S1 during the stimulus presentation window (50, 50 neurons from 4 mice for LN, LE respectively), either before (LN) or after acquisition (LE) of a go/no-go texture discrimination task. Hit trials occur when the reward associated stimulus (s_1) was correctly responded to with the go action. CR trials occur when the unrewarded stimulus (s_2) was correctly responded to with the no-go action. **b**, Summed firing rates of the simulated sensory neurons, $\sum_i z_i$, depending on the sensory stimulus. Stimulus 1 (s_1) is associated with reward and stimulus 2 (s_2) is not. Statistical differences were tested with two-tailed, independent samples T-tests with Bonferroni corrections.

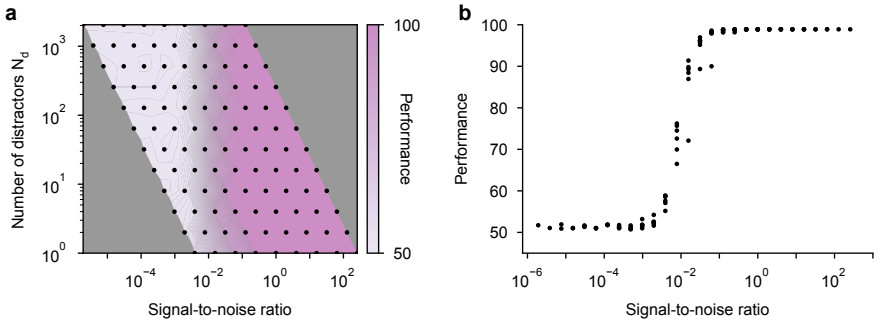

Fig. S3 Signal-to-noise ratio determines learning. a, Final performance for the model (see panel a) without gain modulation depending on the number of distracting neurons (N_d) and the signal-to-noise ratio (SNR), see equation (13). **b**, From panel b, we suspect that the signal-to-noise ration (SNR) makes a good prediction for the final performance of the model without gain modulation. This can be emphasized by plotting the performances, but only with regard to the signal-to-noise ratio. Each point corresponds to a point from panel a.

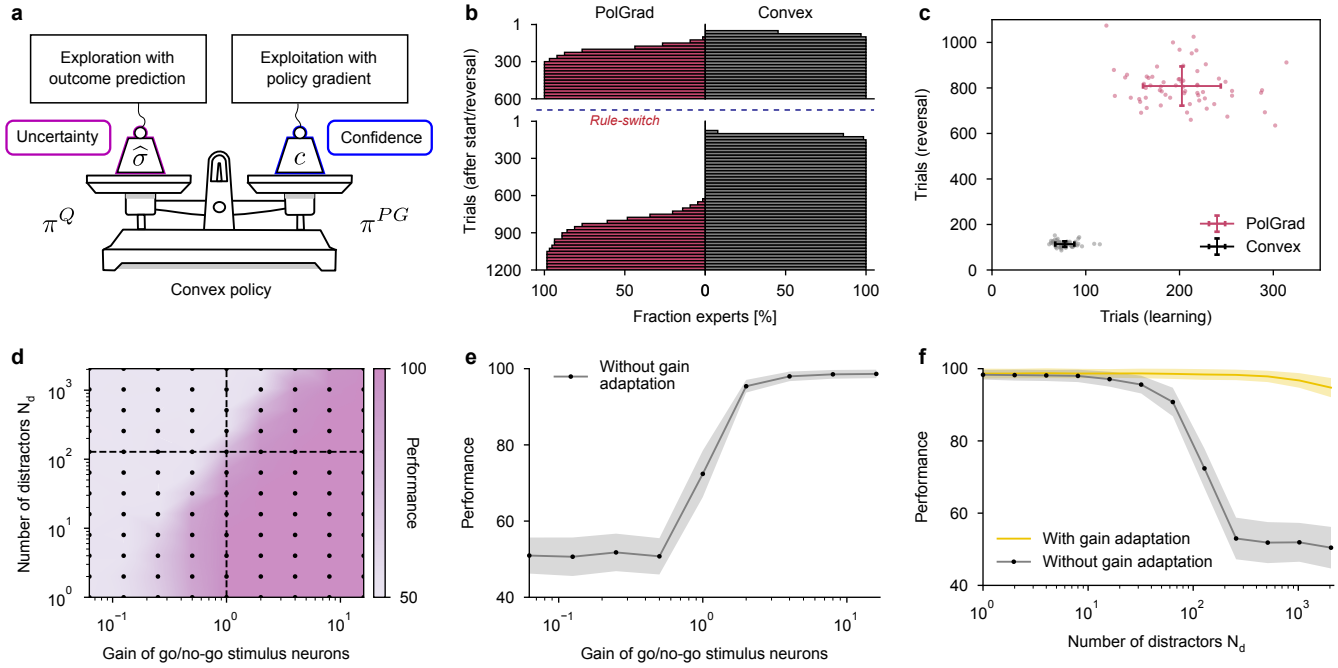

Fig. S4 Convex policy for action selection in non-stationary environment. a, Schematic of the convex action selection policy, which weighs between an exploitation policy and an exploration policy according to the prediction confidence. **b**, Cumulative distributions of number of trials necessary to reach expert performance for either the original and the reversed rule. This evolution is plotted for the policy gradient method (PolGrad) and for the convex policy (Convex) in both cases without any distractors. In both cases the parameters were optimized for the final performance after the rule reversal. **c**, Two-dimensional comparison of the time to reach expertise for the policy gradient and for the convex method. The convex method is adapts more rapidly and more consistently. **d**, **e** and **f** show the same as Fig. 1b, 1c and 1f for a convex action selection policy instead of pure policy gradient. This confirms that the dependence on a high signal-to-noise ratio and the advantage of task-dependent gain modulation also applies for the convex policy.

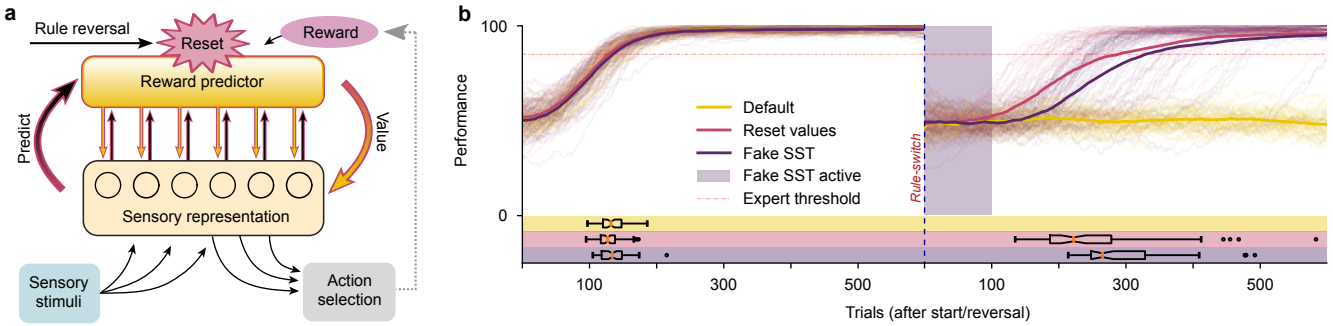

Fig. S5 Artificial reset of top-down modulation upon rule reversal. a, The 'Reset values' model helps us study the importance of top-down gain adaptation during reversal learning by artificially resetting the apical modulation apparatus after a rule change (see Fig. 1e for comparison). Upon rule reversal, the reward predictor and the top-down apical synapses are reinitialized and restart learning from scratch, which negates any bias produced by learning from the previous rule. **b**, Comparing the performance traces of different gain adaptation algorithms shows that apical dendrite inhibition after a rule reversal can help unbias the sensory representation and support reversal learning. The default gain modulation model without lOFC (yellow) struggles to learn after reversal. The 'Reset values' model (red) performs much better, since it does not suffer from ill-suited top-down amplification after the reversal. Finally, the 'Fake SST' simulation (violet) was tested, which took the default model without lOFC and increased the SST interneuron activity to a high value (10) for a duration of 100 trials after the rule reversal. The fact that the 'Fake SST' model adapted almost as well as the artifical 'Reset values' model shows that well timed apical top-down inhibition via SST interneurons can provide a biological mechanism to unbias the sensory top-down amplification.

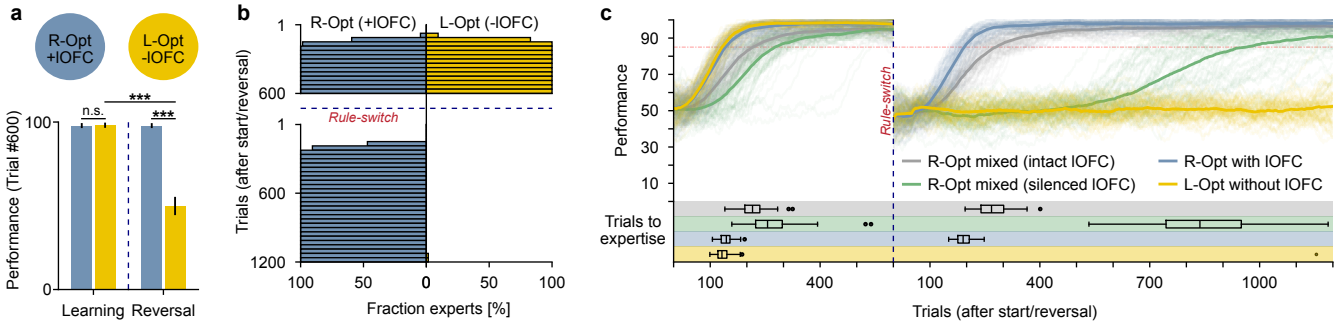

Fig. S6 Alternative parameter optimization. a, The main model parameters (used in the main figures) are derived to be the same in the simulations where the lOFC is intact or silenced (see SI). For completeness purposes, here we also show the performance of the model without lOFC (-lOFC), where the parameters are optimized for performance in the original rule (L-Opt). These are compared to the model with lOFC (+lOFC), where all parameters are optimized for performance in the reversed rule (R-Opt). **b**, The cumulative distribution of expert agents depending on the trial before and after reversal, again for the model without lOFC optimized for the first learning phase, L-Opt (-lOFC) and the same for the model with lOFC optimized for the reversal phase, R-Opt (+lOFC). **c**, Comparison of the performance traces the model without lOFC optimized for the learning phase (L-Opt without lOFC, yellow) and the model with lOFC optimized for the reversal phase (R-Opt with lOFC, blue) with the model optimized jointly for the reversal phase (R-Opt mixed) either with the lOFC intact (green) or silenced (grey).

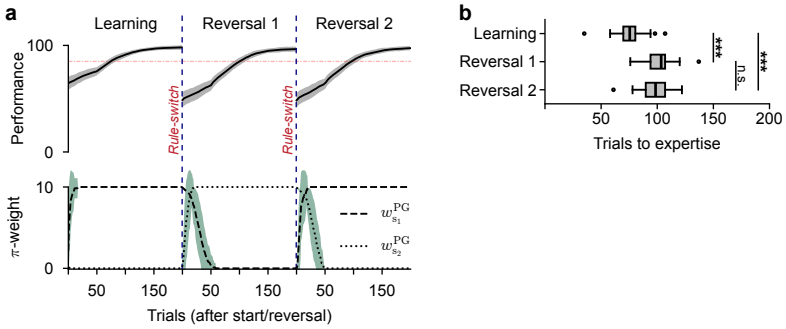

Fig. S7 Catastrophic forgetting without gain modulation. a, Evolution of the performance trace (top) and π-weights (bottom) without gain modulation and without distracting stimuli. The policy weights are unlearned after each reversal in order to allow the behavior to adapt to the new rule. **b**, Number of trials necessary to reach expert performance (85%) during the first learning phase (Learning), after one rule reversal (Reversal 1) and after the second reversal (Reversal 2). Statistical differences were tested with two-tailed, independent samples T-tests with Bonferroni corrections.

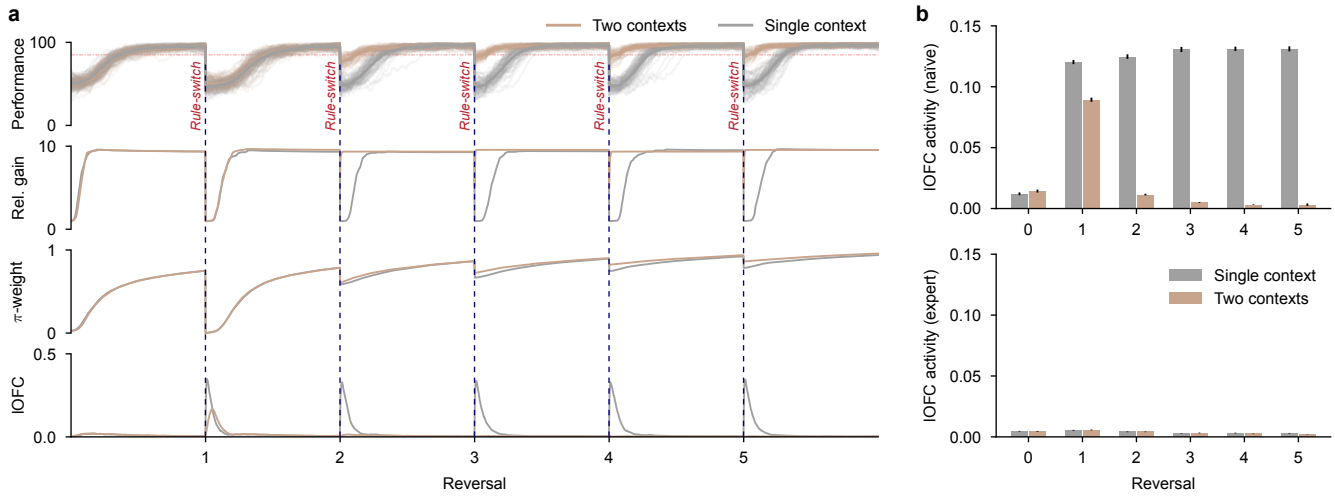

Fig. S8 lOFC as context switch. a, Average performance trace during five sequential rule reversals (top row) as well as the evolution of the average relative gain and the average π-weight for the go-stimulus neuron of the current rule (middle rows). The rapid adaptation of the relative gain to the go-stimulus explains the increased adaptation speed shown in Fig. 4g. It also shows the evolution of the lOFC activity (bottom) modeled by the context-prediction error (CPE). All the mean traces are shown both for the model with a single context representation (grey) and with two context representations (brown). **b**, Average lOFC activity (modeled as CPE) either during the naïve phase (first 100 trials) or the expert phase (last 100 trials) of each reversal cycle (of 600 trials each). The error bars show the standard error.

Table 1 Model combinations

Action-selection policy	Gain modulation	Figure occurrence
policy gradient	none	1efh, S3, S4bc
policy gradient	reward-guided amplification (without lOFC)	1hi
fixed policy	reward-guided amplification (without lOFC)	2cf, S1d
convex policy	none	S4, S7
convex policy	reward-guided amplification (without lOFC)	2b, 3, S4f, S5b, S6
convex policy	reward-guided amplification + lOFC targets SST	3, 4, 5, S2, S6, S8
convex policy	reward-guided amplification + lOFC switches context representation	4g, S8

Table 2 Fixed parameter values

Parameter description	Mathematical symbol	Parameter value
Maximal apical gain	$g_{\max}$	10
Reward value	R_{reward}	1
Punishment value	$R_{\text{punishment}}$	-1
L1 regularization factor	λ^V	$1/N$
Excitatory apical weight learning rate	η^{ap}	0.1
Maximal excitatory apical weight	$w^{\text{ap}}_{\max}$	$g_{\max} - 1 = 9$
Minimal excitatory apical weight	$w^{\text{ap}}_{\min}$	0
Uncertainty learning rate	η^{σ}	0.03
lOFC learning rate	η^{lOFC}	0.03
SST interneuron baseline activity	r_0^{SST}	0.1
Q-value policy temperature	T^{Q}	10
Minimal policy exploration	ϵ_ϕ	0.01
Policy sigmoid bias	b_ϕ	99
Maximal uncertainty	$\sigma^2_{\max}$	0.5
CPE threshold for context switch	θ_{switch}	1
Confidence threshold for context switch	c_{switch}	0.95

Table 3 Optimized parameter values

Model description	η^π	η^V	η^Q	r_0^{VIP}	w^{SST}	Optimized for	Figure occurrence
Policy gradient without gain	M	N/A	N/A	N/A	N/A	Learning perf.	1efh, S3
Policy gradient with gain (without lOFC)	M	M	N/A	N/A	N/A	Learning perf.	1hi
Fixed policy with gain (without lOFC)	N/A	$10^{1.25\dagger}$	$10^{-1.5\dagger}$	N/A	N/A	Reversal time	2cf, S1d
Convex policy with gain (without lOFC)	$10^{-2.5}$	$10^{0.75}$	$10^{-1.5}$	N/A	N/A	Reversal time	3, S6
Convex policy with gain (with lOFC)	$10^{-2.5\dagger}$	$10^{1.25}$	$10^{-1.5\dagger}$	$10^{-0.75}$	$10^{1.5}$	Reversal time	3, 4, 5, S2, S6, S8
Policy gradient without gain (no noise)	N/A	10^{-1}	N/A	N/A	N/A	Reversal time	S4bc
Convex policy without gain (no noise)	$10^{1.5}$	N/A	$10^{1.5}$	N/A	N/A	Reversal time	S4bc, S7
Convex policy without gain	M	N/A	M	N/A	N/A	Learning perf.	S4def
Convex policy with gain (without lOFC)	M	M	M	N/A	N/A	Learning perf.	2b, S4f
Convex policy with gain (without lOFC)	$10^{-1.5}$	$10^{1.25}$	$10^{1.5}$	N/A	N/A	Learning time	S5, S6

N/A stands for no value available because the parameter is not applicable for the model in question
M stands for multiple values available, because the optimization was run multiple times (e.g. for different numbers of distractors)
† means that the parameter value was copied from another parameter optimization procedure.

Algorithm 1 Simulation of one trial

function SIMULATETRIAL($T, \boldsymbol{x}^{\text{bas}}, \boldsymbol{x}^{c}, \boldsymbol{w}^{\text{ap}}, \pi, \hat{V}, \hat{Q}, \hat{\sigma}, \text{CPE}$)
- $\boldsymbol{x}^{\text{bas}} \leftarrow T.getInput()$ ▷ Get the bottom-up sensory inputs for the current trial T
- $r^{\text{SST}} \leftarrow getSST(\text{CPE})$ ▷ Compute SST interneuron activity, equation (10)
- $\boldsymbol{z} \leftarrow x^{\text{bas}} \odot \boldsymbol{g}(\boldsymbol{w}^{\text{ap}} x^{c} - r^{\text{SST}})$ ▷ Compute sensory representations, equations (1, 2)
- $\boldsymbol{w}^{\text{ap}}.update(\hat{V}, \boldsymbol{z})$ ▷ Update excitatory top-down weights, equation (5
- $a \leftarrow \pi(\boldsymbol{z})$ ▷ Decide the action to take, equation (11)
- $\Omega \leftarrow T.Outcome(a)$ ▷ Observe the outcome for the trial T given action a
- $R \leftarrow T.Reward(\Omega)$ ▷ Observe the reward for the trial T given action a
- $\hat{V}.update(R)$ ▷ Update state value estimator, equation (4)
- $\pi.update(R)$ ▷ Update policy network, equation (21)
- $\text{CL} \leftarrow getCL(\hat{\sigma}, R - \hat{Q}(\boldsymbol{z}, a))$ ▷ Compute confidence loss, equation (8)
- $\text{CPE}.update(\text{CL})$ ▷ Update CPE, equation (9)
- $\hat{\sigma}.update(R - \hat{Q}(\boldsymbol{z}, a))$ ▷ Update uncertainty estimate, equation (6)
- $\hat{Q}.update(\Omega)$ ▷ Update state action value estimator, equation (15)
- $\boldsymbol{x}^{c}.update(\text{CPE}, c)$ ▷ Switch context representation (if CPE $> \theta_{\text{switch}}$ and $c > c_{\text{switch}}$)

end function

Algorithm 2 Iterative grid search parameter optimization

function TESTPARAMS($parameters\_sets, best\_loss$) ▷ Find the best set of parameters out of multiple sets
- **for** $p = 0$ to $parameters\_sets.Size() - 1$ **do** ▷ Iterate over all sets of parameters of the grid
  - $parameters \leftarrow parameters\_sets[p]$ ▷ The current set of parameters
  - $loss \leftarrow Simulate(parameters)$ ▷ Simulate the model (64 times)
  - **if** $loss < best\_loss$ **then** ▷ If the loss is smaller than the currently best loss
    - $best\_loss \leftarrow loss$ ▷ Update the best loss
    - $best\_parameters \leftarrow parameters$ ▷ Update the best set of parameters
  - **end if**
- **end for**
- **return** $best\_loss, best\_parameters$

end function

function OPTIMIZEPARAMS
- $iterations \leftarrow 2$ ▷ Number of grid tightening procedures to run
- $parameters\_sets \leftarrow InitialGrid()$ ▷ Get the initial logarithmic parameter grid for the model
- $loss \leftarrow \text{inf}$ ▷ The loss for the best set of parameters
- $loss, best\_parameters \leftarrow$ TESTPARAMETERS($parameters\_sets, loss$) ▷ Find the best set of parameters
- **for** $i = 1$ to $iterations$ **do** ▷ Tighten the resolution of the grid search multiple times
  - $converged \leftarrow False$ ▷ Whether the grid search has converged for a given resolution
  - **while not** $converged$ **do**
    - $parameters\_sets \leftarrow surroundGrid(best\_parameters, i)$ ▷ Get new logarithmic parameter grid
    - $loss, parameters \leftarrow$ TESTPARAMETERS($parameters\_sets, loss$) ▷ Find the best set of parameters
    - **if** $parameters = best\_parameters$ **then** ▷ If the best parameters have not changed
      - $converged \leftarrow True$ ▷ The search has converged for this grid resolution
    - **else**
      - $best\_parameters \leftarrow parameters$ ▷ Update best parameter set
    - **end if**
  - **end while**
- **end for**
- **return** $best\_parameters$

end function