bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Relevance learning via inhibitory plasticity and its implications for schizophrenia

Nathan Insel [1,*] and Blake A. Richards [2,3,4,*]

1 Department of Psychology, University of Montana, 32 Campus Dr. Missoula, MT, 59801 USA
2 Department of Biological Sciences, University of Toronto Scarborough, 1265 Military Trail. Toronto, Ontario M1C 1A4 Canada
3 Department of Cell and Systems Biology Sciences, University of Toronto, 25 Harbord St. Toronto, Ontario M5S 3G5 Canada
4 Learning in Machines and Brains Program, Canadian Institute for Advanced Research

*** Corresponding authors, email: blake.richards@utoronto.ca or nathan.insel@umontana.edu**

Abstract

Symptoms of schizophrenia may arise from a failure of cortical circuits to filter-out irrelevant inputs. Schizophrenia has also been linked to disruptions to cortical inhibitory interneurons, consistent with the possibility that in the normally functioning brain, these cells are in some part responsible for determining which inputs are relevant and which irrelevant. Here, we develop an abstract but biologically plausible neural network model that demonstrates how the cortex may learn to ignore irrelevant inputs through plasticity processes affecting inhibition. The model is based on the proposal that the amount of excitatory output from a cortical circuit encodes expected magnitude of reward or punishment ("relevance"), which can be trained using a temporal difference learning mechanism acting on feed-forward inputs to inhibitory interneurons. The model exhibits learned irrelevance and blocking, which become impaired following disruptions to inhibitory units. When excitatory units are connected to a competitive-learning output layer, the relevance code is capable of modulating learning and activity. Accordingly, the combined network is capable of recapitulating published experimental data linking inhibition in frontal cortex with fear learning and expression. Finally, the model demonstrates how relevance learning can take place in parallel with other types of learning, through plasticity rules involving inhibitory and excitatory components respectively. Altogether, this work offers a theory of how the cortex learns to selectively inhibit inputs, providing insight into how relevance-assignment problems may emerge in schizophrenia.

Introduction

Many symptoms of schizophrenia can be understood as an inability of the brain to appropriately assign relevance to environmental stimuli and internal representations. Schizophrenic patients exhibit difficulties filtering-out, or gating, irrelevant external stimuli (Mcghie and Chapman, 1961; Venables, 1964; Lang and Buss, 1965; McGhie, 1970; Garmezy, 1977; Shagass, 1976; Venables, 1977; Garmezy, 1977), and delusions may also be the product of misattributing relevance (what others have labeled "salience") to certain types of internally-generated representations (Kapur, 2003). Convergent evidence suggests this may result from a dysfunction of inhibitory processes. This idea may have been first proposed by Johnson (1985), who hypothesized that that schizophrenia symptoms arise specifically from a failure of feed-forward inhibition–i.e., activation of inhibition by a system's inputs. Circuits for cortical feedforward inhibition are now relatively well defined, and may principally involve fast-spiking, parvalbumin-expressing (PV+)

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

inhibitory interneurons (Agmon and Connors, 1992; Gibson et al., 1999; Dantzker and Callaway, 2000; Gonchar and Burkhalter, 2003). It is also now well established that PV+ interneurons are compromised in schizophrenia (reviewed by (Benes and Berretta, 2001; Lewis et al., 2005; Lewis and Moghaddam, 2006; Lewis, 2014; Gonzalez-Burgos et al., 2015)).

Computational models have helped to articulate the link between inhibitory dysfunction and schizophrenia (Vogels and Abbott, 2007, 2009; Murray et al., 2014). An important example is work by Vogels and Abbott (2007, 2009) which demonstrated how inhibition may serve to selectively gate some representations but not others. A theme of these models is the importance of balanced excitation and inhibition (EI balance) within the network. EI balance has been extensively studied across a range of cortical regions (e.g., auditory cortex (Wehr and Zador, 2003; Zhang et al., 2003; Dorrn et al., 2010), somatosensory cortex (Gabernet et al., 2005; Wilent and Contreras, 2005; Daw et al., 2007; Chittajallu and Isaac, 2010), olfactory cortex (Poo and Isaacson, 2009), visual cortex (Anderson et al., 2000; Xue et al., 2014), and frontal cortex (Haider et al., 2006)). EI balance may fluctuate dynamically, such as through disinhibition, the importance of which has gained increasing recognition (Pi et al., 2013; Isaacson and Scanziani, 2011; Carcea and Froemke, 2013; Letzkus et al., 2015; Kato et al., 2015). Therefore, a better understanding of the relationship between cortical disinhibition and relevance coding may be critical to understand schizophrenia.

In the present study, we sought to fill gaps in our understanding of the link between relevance assignment and neural inhibition. Two main questions are addressed. First, what may be the relationship between relevance-coding and the dynamic fluctuations in EI balance? Second, how might inhibitory components of the cortical circuit learn the relevance of specific input patterns? Answering these questions will help explain how "gating" is learned, and thereby also help to explain symptoms of schizophrenia. To this end, we have developed an abstract, but biologically plausible, neural network model that is capable of learning to ignore specific inputs, but not if inhibition is disrupted. The fundamental proposal in the model is that the overall level of excitation in a cortical circuit can be interpreted as a signal of the temporally discounted expectation of rewards and/or punishments (**Figure 1A**; see also Insel and Barnes 2015). According to this formulation, deviations in EI balance come to represent the network's estimate of the magnitude of the value signal used in reinforcement learning (Sutton and Barto, 1998).

Three sets of simulations are used to demonstrate the explanatory power of the model. The first demonstrate the model's capacity to learn input relevance/irrelevance, and that, paralleling symptoms of schizophrenia (e.g., (Jones et al., 1992, 1997; Oades et al., 2000)), learning is disrupted by impaired inhibition. The second set of simulations use an extended model to show how the proposed relevance code can modulate learning and, furthermore, reproduce the effects of manipulating inhibition in rodent frontal cortex (Piantadosi and Floresco, 2014; Courtin et al., 2014). The final set of simulations show how relevance learning could occur concurrently with other types of learning, e.g. learning to categorize input patterns, thereby providing a mechanism to multiplex information about stimulus-relevance and stimulus-identity.

Methods

Network summary

The basic structure of the neural network model is illustrated in **Figure 1B**. The core of the model is a spiking, two-layer feedforward neural network composed of different types of units. Stimuli are encoded by a set of input units, the 'Sensory' layer, $\mathbf{x}(t) = [x_1(t), \ldots, x_n(t)]$ (for notation purposes, we use bold symbols for any variable that represents a population of neurons in the model). This layer is a set of excitatory Poisson units with firing rates $\boldsymbol{\psi}(t) = [\psi_1(t), \ldots, \psi_n(t)]$. Changes in the Sensory layer take place when stimuli are presented, as described in more detail below. Sensory units feed into a middle layer, 'Cortex' that is comprised of two populations: an excitatory population of linear-non-linear-Poisson units (Paninski et al., 2007) with rates $\boldsymbol{\lambda}(t) = [\lambda_1(t), \ldots, \lambda_m(t)]$, and spiking activity $\mathbf{E}(t) = [e_1(t), \ldots, e_m(t)]$, and a single inhibitory population unit, $\mathbf{I}(t)$, which receives $\mathbf{x}(t)$ and inhibits the excitatory neurons divisively. The receptive fields of Cortex units have no temporal dimension, so the activity at any point

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

only reflects the current inputs to the network. The inhibitory unit is intended to model the entire population of inhibitory interneurons that provide feed-forward inhibition in the circuit. We justify this simplification in part by noting evidence that feed-forward inhibition in cortex provides a non-specific "blanket" of inhibition to excitatory neurons (Fino et al., 2013; Karnani et al., 2014).

The model is highly abstract, and makes a number of simplifying assumptions for the sake of mathematical tractability. One notable simplification, in comparison to true cortical networks, is the lack of feedback connections between excitatory units. The decision to omit these eliminated the need to explicitly model feedback inhibition to balance recurrent excitation, which helped focus the present investigation on the hypothesis that plasticity in feed-forward inhibition can support relevance learning. A second notable simplification is the absence of explicit signaling loops through other structures (e.g. the basal ganglia, thalamus, or other layers of cortex), which may be necessary for relevance learning in the brain. Naturally, the computations illustrated by the present, simplified network may be implemented in the brain through more steps and across more regions. Ultimately, the model serves as a proof of concept that speaks to only certain elements of true cortical computation.

The connections from input units to the excitatory cortex units are contained in an $m \times n$ synaptic weight matrix, $W^{\mathbf{x}\to\mathbf{E}}$, the connections from input units to the inhibitory Cortex unit are contained in the n-dimensional vector of synaptic weights, $W^{\mathbf{x}\to\mathbf{I}}$, and the connections from the inhibitory unit to the excitatory units are contained in an m-dimensional vector of synaptic weights, $W^{\mathbf{I}\to\mathbf{E}}$.

Altogether, this set-up gives the following equations which describe the activity of the model:

$$\begin{aligned}
\mathbf{x}(t) &\sim Poisson(dt\boldsymbol{\psi}(t)) \\
\mathbf{I}(t) &= W^{\mathbf{x}\to\mathbf{I}}\mathbf{x}(t) + b^{\mathbf{I}} \\
\boldsymbol{\lambda}(t) &= \frac{W^{\mathbf{x}\to\mathbf{E}}\mathbf{x}(t) + b^{\mathbf{E}}}{W^{\mathbf{I}\to\mathbf{E}}\mathbf{I}(t)} \\
\mathbf{E}(t) &\sim Poisson(dt\boldsymbol{\lambda}(t))
\end{aligned} \tag{1}$$

where dt is the time-step, which is 20 ms for most simulations, and $b^{\mathbf{E}}$ and $b^{\mathbf{I}}$ are bias terms.

One additional component that is not included in the above equations, but which contributes to relevance learning (see Relevance Learning, below), is a signal communicating the unsigned magnitude of reward or punishment, i.e. the unconditioned stimulus (US). In the present simulations the value of the US at a given time ($u(t)$) is either 1 or 0, though in principle it could as easily be a graded value with 1 representing the highest degree of relative salience and 0 the lowest.

In some simulations, we add an additional output layer of units with activity $\mathbf{y}(t) = [y_1(t), \ldots, y_\ell(t)]$ that receives inputs from the excitatory cortical units via an $\ell \times m$ synaptic weight matrix, $W^{\mathbf{E}\to\mathbf{y}}$. In those simulations which address the effects of frontal manipulations on fear behavior, the output layer is intended to represent an amygdala ('Amygdala') and implements a competitive learning algorithm (according to the framework of Rumelhart and Zipser 1985). In the competitive learning module, a maximum of only one unit may be active at any given time (it is possible for no units to be active). Whether a unit, i, is active depends on two conditions: (1) the unit is receiving stronger input than any of the other units, (2) the unit's input, $z_i(t)$ is greater than a threshold θ. Amygdala units also receive signals from the US, such that $u(t)$ can help to increase input, $z_i(t)$. Based on all of this, the activities of the Amygdala units are governed by the following equations:

$$\begin{aligned}
z_i(t) &= W_i^{\mathbf{E}\to\mathbf{y}}\mathbf{E}(t) + u(t)\theta \\
\mathbf{y}(t) &= \begin{cases} z_i(t) & \text{if } z_i(t) > z_j(t),\ \forall j \neq i \text{ and } z_i(t) \geq \theta \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{2}$$

where $W_i^{\mathbf{E}\to\mathbf{y}}$ is the vector of synaptic weights into the i_{th} Amygdala unit. There are some important aspects of this to note:

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1. All units will pass threshold, θ, if $u(t) = 1$.
2. Every $W_i^{\mathbf{E}\to\mathbf{y}}$ vector has its magnitude normalized such that $||W_i^{\mathbf{E}\to\mathbf{y}}||_2 = 1$ (similar to Rumelhart and Zipser 1985).
3. Only the unit whose weight vector $W_i^{\mathbf{E}\to\mathbf{y}}$ is most closely aligned to $\mathbf{E}(t)$ (according to the dot product) will be active.

The threshold, θ, determines when the Amygdala layer can have any active units in the absence of an unconditioned stimulus (i.e. if $u(t) = 0$). An explanation for how θ was derived mathematically is given below in the Amygdala learning section.

Although having a single neuron firing is undoubtedly not what occurs in the mammalian amygdala, there is evidence for a competitive "winner-takes-all" mechanism (Han et al., 2007; Rashid et al., 2016), such that a single ensemble of neurons is active and all others are silent. Therefore, the active unit in our model 'Amygdala' could be taken to represent an ensemble of "winning" neurons. Since individual units in this case were representative of larger ensembles, the winning unit's firing rate was kept as a continuous "activation level" value rather than Poisson-distributed spiking.

The output layer takes on a different form in those simulations that demonstrate how our model can multiplex relevance signals and stimulus identity. In this case, the output units represent some efferent, such as a second area of cortex, that is responsible for categorizing input activity. For simplicity, we refer to this layer in the simultations as the 'Category' layer. The Category layer is a set of softmax, linear-non-linear-Poisson units with rates $\boldsymbol{\phi}(t)$ governed by:

$$\begin{aligned} \phi_i(t) &= 20\frac{W_i^{\mathbf{E}\to\mathbf{y}}\mathbf{E}(t)}{\sum_j W_j^{\mathbf{E}\to\mathbf{y}}\mathbf{E}(t)} \\ \mathbf{y}(t) &\sim Poisson(dt\boldsymbol{\phi}(t)) \end{aligned} \tag{3}$$

where the activity of the units is proportional to the probability of each of the ℓ possible categories for the current stimulus (a rate-of-fire of 20 Hz corresponds to a probability of 1).

Stimuli

Sensory (input) units are divided into sets of stimulus-coding and non-coding units. Each stimulus is capable of activating one tenth of the Sensory units. In the case of the Learned Irrelevance simulation, for example (see Learned irrelevance and blocking, below), one tenth of the units are activated by the stimulus, "CS+" that is paired in time with the US, one tenth of the units are activated by another stimulus, "CS-", that is not paired with the US, and the rest are activated by neither CS+ nor CS-. Importantly, Sensory units fire both when activated by a stimulus and not, just at different rates. An active sensory unit fires spikes with a Poisson process at a rate of $\psi^{on} = 20$ Hz, while an inactive sensory unit fires spikes with a Poisson process at a rate selected randomly based on a gamma distribution that peaks at 0.6 Hz and has a variance of 3 Hz. These rates were selected based on baseline firing characteristics among putative excitatory neurons recorded from the rat medial prefrontal cortex (Insel and Barnes, 2015). Hence, for example, if the CS+ is presented to the network then the ten percent of the units activated by CS+ will be firing at a rate of 20 Hz, while other units will continue firing at their typically low (0-2 Hz) but occasionally high (10 or 20 Hz) baseline rate. In those simulations that use the Category output layer, Sensory units are divided into ten sets, with each set activated differentially by a particular category (in these simulations, baseline rates were also simplified to be homogeneously 2 Hz, as variance was found to not impact the results).

Weights

Initialization of the three sets of connection weights–Sensory to Cortex excitatory units ($W^{\mathbf{x}\to\mathbf{E}}$), Sensory to Cortex inhibitory unit ($W^{\mathbf{x}\to\mathbf{I}}$), and Cortex inhibitory to excitatory units ($W^{\mathbf{I}\to\mathbf{E}}$)–took into account

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

three issues. First, when novel stimuli were first presented to the network, the evoked activity in Cortex excitatory units needed to be higher than baseline levels, but not higher than levels associated with "relevant" stimuli (described below in Relevance learning). Second, baseline activity of the cortex inhibitory unit had to be high enough that reducing $W^{\mathbf{x}\to\mathbf{I}}$ had an impact on Cortex excitatory units. Third, that $W^{\mathbf{x}\to\mathbf{I}}$ were balanced with $W^{\mathbf{x}\to\mathbf{E}}$, such that small changes in $W^{\mathbf{x}\to\mathbf{I}}$ could not dramatically alter population activity.

Starting weights were assigned as follows: all inhibitory to excitatory weights ($W^{\mathbf{I}\to\mathbf{E}}$) were set to 0.4. Sensory to inhibitory weights ($W^{\mathbf{x}\to\mathbf{I}}$) were set to 0.5, although in most simulations these were adjusted during pre-exposure periods to establish a desired level of Cortex activity. In the initial simulations which did not make use of an output layer, Sensory to Cortex excitatory unit weights ($W^{\mathbf{x}\to\mathbf{E}}$) were selected randomly from a Gaussian distribution centered at 0.1, with a standard deviation of 0.4, and any negative connections were clipped at zero. This yielded an average weight of 0.2, with many at zero. In those simulations which used the Amygdala output layer, fear learning to the CS+ required that Sensory activity patterns could be interpreted from Cortex excitatory unit patterns; therefore, Sensory to Cortex excitatory weights ($W^{\mathbf{x}\to\mathbf{E}}$) were set to a Gaussian-smoothed identity matrix (the smoothing window was equal to the number of inputs divided by 20; e.g., for 500 units, a window of 25 was used), then multiplied by 1/3 the number of inputs. In those simulations which used the Category output layer, Sensory to Cortex excitatory weights were modified according to stimulus presentations; initialization weights were therefore arbitrarily drawn from a uniform distribution over $[0-1]$.

Relevance Learning

The principal learning mechanism used in this paper is a modification of the temporal difference learning algorithm (Sutton and Barto, 1998). Specifically, a population-based relevance (or "salience" signal), $S(t)$ reflects the deviations in excitatory activity from an established baseline. The baseline level can be thought of as the EI balance set-point maintained by the cortical network. The level of Cortex excitatory unit activity was measured as the vector norm of the population of spiking excitatory units, $||\mathbf{E}(t)||_2 = \sqrt{\sum_i e_i(t)^2}$ (the reasons for using the norm become clear in the following section on Amygdala learning). $S(t)$ is therefore determined by the difference between $||\mathbf{E}(t)||_2$ and the homeostatic set-point for the population, H (**Figure 1B**):

$$S(t) = ||\mathbf{E}(t)||_2 - H \tag{4}$$

The goal of relevance learning in the network is to have $S(t)$ come to represent expected relevance, which we interpret as "unsigned value", $U(t)$:

$$U(t) = \langle \sum_{i=1}^{\infty} \gamma^{i-1} u(t+i) \rangle \tag{5}$$

where $u(t)$ is the unsigned reward/punishment signal, US, described above, γ is a temporal discounting factor, and $\langle\cdot\rangle$ indicates expected value. $U(t)$ is akin to the value function, $V(t)$, used in temporal difference learning (Sutton and Barto, 1998). Similar to temporal difference learning, the goal of learning in our model is, in part, to ensure that $S(t)$ is a good estimate of $U(t)$. This is accomplished using a prediction error signal, $\beta(t)$:

$$\beta(t) = H^* u(t) + \gamma S(t) - S(t-1) \tag{6}$$

where H^* is a salience scaling factor that determines how much cortical activity levels should deviate from the set point in response to relevant stimuli. We use β to represent our prediction error signal, rather than the usual δ, to distinguish it from prediction error signals that measure differences in signed

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(as opposed to unsigned) value estimates (Sutton and Barto 1998; the notation also deviates slightly from convention by using "t" rather than "t+1", to avoid questions about whether the model has future information).

To put it another way, the system learns to ensure that fluctuations in $||\mathbf{E}(t)||_2$ away from the set-point, H, reflect experience with rewards/punishments (the US). The scale of the fluctuations is determined by H^*. Training the salience signal $S(t)$ involves updating the synaptic weights in Cortex to achieve $\beta(t) = 0$. It can be seen that this is achieved when:

$$||\mathbf{E}(t)||_2 = H + H^* U(t) \implies S(t) = H^* U(t) \tag{7}$$

since $U(t-1) = \langle u(t) \rangle + \gamma U(t)$. Therefore, β is generally close to zero when the following conditions have been achieved: (1) for stimuli that do not predict any reward or punishment, the norm of the spike count in the excitatory cortical population is equal to the homeostatic constant, H and (2) for stimuli that do predict reward or punishment $S(t)$ is a linear function of $U(t)$ with a slope of H^*. Once again, the highest possible US ($u(t)$) value is 1 based on the assumption that natural stimuli, no matter how salient, should not be capable of pushing network activity beyond a limit.

To learn this, we apply a Hebbian weight update modulated by the prediction error, $\beta(t)$. In most simulations the weight change takes place on the feed-forward connections to the inhibitory unit, described by the following update equation:

$$\Delta W_i^{\mathbf{x}\to\mathbf{I}}(t) = -\alpha\beta(t) x_i(t) \mathbf{I}(t) \tag{8}$$

where Δ indicates a weight update, i indexes sensory units, α is the learning rate, and the negative sign is required because the weight change is on synapses that have an overall inhibitory effect. The synapses were updated on-line at each time-step of the simulations.

In early simulations, the performance of this learning rule is compared against modified rules in which either the $W^{\mathbf{x}\to\mathbf{E}}$ or $W^{\mathbf{I}\to\mathbf{E}}$ synapses are modified:

$$\begin{aligned} \Delta W_{ij}^{\mathbf{x}\to\mathbf{E}}(t) &= \alpha\beta(t) x_j(t) e_i(t) \\ \Delta W_i^{\mathbf{I}\to\mathbf{E}}(t) &= -\alpha\beta(t) \mathbf{I}(t) e_i(t) \end{aligned} \tag{9}$$

where j refers to cortex excitatory units.

An important question raised by the above relevance-learning equations is how real cortical networks implement a homeostatic set-point, and what may account for its deviation during rewards/punishers. We offer two non-exclusive possible conceptualizations. The first follows from a tradition of research on prediction error signals in neuromodulatory systems like dopamine and norepinephrine, the second builds off of recent research on the mechanisms by which EI balance is maintained within cortical networks. In the former scenario, the salience signal is explicitly read out by cells in neuromodulatory nuclei, as has been described in dopaminergic nuclei (Bromberg-Martin et al., 2010), which is then used to compute a prediction-error signal that feeds back to the cortical afferent, modifying plasticity (the Hebbian weight update) accordingly. In the second scenario, the prediction error calculation responsible for maintaining EI balance are local to the cortex, and may take place, for example, by intrinsic signaling processes within inhibitory neurons in response to feedback from local excitatory neurons (i.e. the salience signal, see Vogels et al. 2011; Luz and Shamir 2012). Either the salience signal itself (in this scenario, the inhibitory neuron responses to excitatory signals) or the cellular processes responsible for computing the difference between those signals and the desired level (the prediction error signals) are modified by neuromodulatory signals carrying information about current rewards/punishers. The present model is built at one level of abstraction above this implementation, remaining agnostic as to where and how the signals are implemented. These details are, however, likely to be valuable for future empirical and theoretical investigations (this is also considered in Discussion).

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Amygdala learning

In simulations with output units, $\mathbf{y}(t)$, such as the Amygdala, synapses between Cortex excitatory units and the output units were also trained. In simulations using an Amygdala output layer (which were used to simulate the effects of frontal cortex manipulations on fear responses), the Cortex-to-Amygdala weights, $W_{ij}^{\mathbf{E}\to\mathbf{y}}(t)$, were trained with a competitive learning algorithm based on Rumelhart and Zipser (1985). More precisely:

$$\Delta W_{ij}^{\mathbf{E}\to\mathbf{y}}(t) = \begin{cases} \alpha_y[x_j(t)/||\mathbf{x}(t)||_2 - W_{ij}^{\mathbf{E}\to\mathbf{y}}] & \text{if unit } i \text{“won”} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where α_y is the learning rate. These weight updates took place simultaneously with weight updates within Cortex for relevance learning.

Importantly, if a given unit in the Amygdala, i, always tends to “win” when presented with excitatory population vectors $\mathbf{E}'(t) = [e_1'(t), ..., e_m'(t)]$ sampled from the rate vector, $\boldsymbol{\lambda}' = [\lambda_1', ..., \lambda_m']$, then this learning algorithm will converge (on average), when:

$$W_i^{\mathbf{E}\to\mathbf{y}} = \frac{\boldsymbol{\lambda}'}{||\boldsymbol{\lambda}'||_2} \tag{11}$$

This implies that the weight vector $W_i^{\mathbf{E}\to\mathbf{y}}$ is actually just $\boldsymbol{\lambda}'$ scaled to have a norm of 1. As such, if the excitatory Cortex units generate activity vectors with consistent norms that scale based on previous associations with reward or punishment (which should be the case if relevance learning has converged, see equation (8)), then when Amygdala learning has converged the input to unit i, $z_i(t)$ (as determined by 2), will be the maximum in response to $\boldsymbol{\lambda}'$. (That is because the dot product of two vectors is a function of the angle between them, with angles of zero giving the greatest dot product when the magnitude is constant). As such, a suitable threshold, θ, can be found to ensure that in the *absence* of a US the Amygdala only responds to stimuli that have been paired with reward or punishment in the past so that the Amygdala learning algorithm could converge. In the simulations presented here, the value of θ was set by grid search so that the probability of any neuron crossing threshold would be very low if no learning had occurred, and very high if a US was present or learning had converged.

Categorization learning

In simulations where we trained the output units to categorize input stimuli, we used back-propagation of error (Rumelhart et al., 1986) to train the weight matrices $W^{\mathbf{x}\to\mathbf{E}}$ and $W^{\mathbf{E}\to\mathbf{y}}$. More precisely, target vectors, $\mathbf{o}(\mathbf{x}(t))$ are defined, where each stimulus provided to Sensory units has a corresponding target vector for the output Category units. The cross-entropy (Rumelhart et al., 1986) between the Category activity and target vectors was used as the loss function to train the network:

$$\mathcal{L}(x(t)) = \sum_{i=1}^{\ell} o_i(\mathbf{x}(t)) \ln(y_i(t)) \tag{12}$$

where $o_i(\mathbf{x}(t))$ is the “target” response to input $\mathbf{x}(t)$ for output unit i, i.e. $o_i(\mathbf{x}(t)) = 1$ if i is the correct category for $\mathbf{x}(t)$, and it is zero otherwise.

For any weight W_{ij} in $W^{\mathbf{x}\to\mathbf{E}}$ or $W^{\mathbf{E}\to\mathbf{y}}$, the weight update is determined by the partial derivative of this loss function with respect to the weight:

$$\Delta W_{ij} = \alpha_y \frac{\partial \mathcal{L}(\mathbf{x}(t))}{\partial W_{ij}} \tag{13}$$

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

where α_y is the learning rate. This ensures that the network learns to correctly categorize the stimuli (i.e., the pattern of Sensory unit activity, $\mathbf{x}(t)$) using the output, Category units $\mathbf{y}(t)$. As with Amygdala learning, the categorization learning proceeded in tandem with the salience learning.

Learned irrelevance and blocking

The first set of simulations tested learned irrelevance and, in a separate protocol, blocking. These were both run using $dt = 20$ ms. At the beginning of the simulations, a 60 s adaptation period without stimulus presentations was simulated, which allowed weights to adjust to the randomly-selected baseline input activity levels. All pharmacological simulations were implemented after adaptation.

Learned irrelevance: Each of two stimuli used an independent, inter-trial-interval jittered between 20 and 30 seconds. Experimental stimuli were simulated as 200 ms periods during which the firing rate of a pre-determined set of Sensory units (10%) was raised to ψ^{on}. One of the two stimuli, CS+, was paired with a reinforcer, or unconditioned stimulus (US) by setting $u(t) = 1$ (see the section Stimuli above). The onset of the CS+ preceded the US by 100 ms. The network was capable of learning the relevance of a temporally offset CS+ because $\beta(t)$ integrated signals across timesteps with a discounting factor.

Blocking: The blocking protocol used the same parameters as that of learned irrelevance, with the exception that there were four phases of stimulus exposures: 1) a pre-exposure phase, in which both CSs (named CS-A and CS-B) were presented 50 times *without* the US (inter-trial-intervals were decreased to 10-15 s to reduce runtime), 2) a conditioning phase, in which CS-A was presented 50 times paired with a US, 3) a blocking phase, in which the CS-B was presented 50 times simultaneously with the CS-A, paired with a US, 4) a test phase, in which the CS-A and CS-B were presented independently 10 times in the absence of a US. To reduce runtimes, the inter-stimulus intervals were also reduced to between 10 and 15 seconds.

Tests of inhibitory disruptions in different model versions: The effects of inhibitory connection strength changes on learned irrelevance and blocking were assessed using different model versions. The model versions differed with respect to which synapses were plastic: $W^{\mathbf{x}\to\mathbf{I}}$, $W^{\mathbf{x}\to\mathbf{E}}$, or $W^{\mathbf{I}\to\mathbf{E}}$, equations (8) and (9) above. The purpose of the test was to evaluate whether disruptions to inhibitory systems correspondingly disrupt learned irrelevance and blocking, as has been hypothesized to take place in schizophrenia. Inhibitory disruptions were made by reducing the degree to which excitatory units could respond to the inhibitory units by 10%:

$$\boldsymbol{\lambda}(t) = 0.9\frac{W^{\mathbf{x}\to\mathbf{E}}\mathbf{x}(t) + b^E}{W^{\mathbf{I}\to\mathbf{E}}\mathbf{I}(t)} + 0.1(W^{\mathbf{x}\to\mathbf{E}}\mathbf{x}(t) + b^E) \tag{14}$$

$$\mathbf{E}(t) \sim Poisson(dt\boldsymbol{\lambda}(t)) \tag{15}$$

Simulation of Experimental Data: Pharmacological Effects on Latent inhibition

Piantadosi and Floresco (2014) demonstrate the effect of GABA-A antagonists on latent inhibition. Latent inhibition refers to the classic behavioral phenomenon whereby it is harder to associate a familiar stimulus (one that a subject has been pre-exposed to) with a reinforcer (Lubow and Moore, 1959). Latent inhibition is also known to be disrupted in schizophrenia (Lubow et al., 7 06; Baruch et al., 1988; Rascle et al., 2001; Gray and Snowden, 2005). As shown in **Figure 4A-D**, a protocol was created that matched the one used in rats (see also McAllister 1997; Enomoto et al. 2011). For processing time purposes, the stimulus and inter-stimulus times used in the original were reduced by a factor of 5. The protocol began with a 60 s adaptation period, followed by three phases: 1) a pre-exposure phase, in which the network was presented with the conditioning stimulus (CS) 30 times (10% of input units, 6 s long, inter-stimulus interval of 6 s), 2) a conditioning phase, in which the CS was presented simultaneously with foot shock ($u(t) = 1$), and 3) a test phase, in which the CS was presented by itself 4 times. The protocol was performed on three pairs of network models, with each pair including one network given pre-exposures and one that was not given pre-exposures. The three pairs simulated the treatment groups

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

used in the original study: animals treated with saline were simulated without any modification to the network, treatment with GABA-A antagonist during conditioning were simulated using a 20% reduction in inhibition, according to equation (14) during the conditioning phase, and treatment with antagonist during testing were simulated with the same disruption during the testing phase. Conditioned fear responses were measured as the maximal response of amygdala units, averaged across all timesteps during CS presentation.

Simulation of Experimental Data: Optogenetic Effects on Fear Expression

Recent work by Courtin et al. 2014 found that inhibition of parvalbumin, fast-spiking neurons in the mouse mPFC can evoke fear responses, while excitation of the same neurons can decrease fear responses. The protocol used in that study was presently simulated as precisely as possible (**Figure 4E-H**), using all of the same parameters as used in the latent inhibition design. To stimulate optical stimulations, 20% reduction in inhibition, according to equation (14). During the pre-conditioning phase, this reduction in the inputs was applied for 250 ms intervals separated by 860 ms (equivalent to 0.9 Hz stimulations, as in the original study). This was followed by a conditioning phase, in which a 6 s CS+ was paired with footshock (i.e., the firing rate of input units coding for the CS was set to ψ^{on} and $u(t) = 1$). As in the previous protocol, all stimulus and inter-stimulus times were decreased from the original study by a factor of 5. One change from the original protocol is that the 1 s US presentation used in the original study was lengthened to the entire CS period. We justify this change based on an assumed difference between real brains and the model: whereas in the brain, activity and plasticity are likely regulated by change, such as the onset or offset of a stimulus, the model treats each time point equivalently. Thus, the period during which the CS is on but US is off will extinguish the associations learned during their concurrence. The CS-US pairings were presented 12 times with an inter-trial interval of 4-30 s. The conditioning phase was followed by an extinction phase, in which the CS was again presented 12 times with the same inter-trial interval, followed in turn by a series of CS presentations accompanying the 40% reduction in $W^{\mathbf{x}\to\mathbf{I}}$. To test the effect of inhibitory activation during a conditioned CS, the same conditioning protocol was used, but was followed by presentations of the CS accompanying increases to inhibitory unit activit. We found that only a 10% increase in $W^{\mathbf{I}\to\mathbf{E}}$ was necessary to elicit changes approximating those observed in the original study.

Pairing of relevance-learning with classification learning

To examine the ability of the network to carry both the salience signal and the other information simultaneously (i.e. to multiplex the salience signal with other signals) simulations were run wherein the feed-forward excitatory weights ($W^{\mathbf{x}\to\mathbf{E}}$ and $W^{\mathbf{E}\to\mathbf{y}}$) were trained to perform categorization of the inputs, $\mathbf{x}(t)$, while the excitatory weights onto the inhibitory unit, $W^{\mathbf{x}\to\mathbf{I}}$, were trained according to the relevance algorithm described in equation (8). To do this, stimuli were presented as described above for 1000 timesteps with $dt = 20$ ms, and the learning proceeded at each timestep as described in Categorization learning, above. For simplicity, however, the network was trained using a batch method, wherein all the weight updates in response to all of the stimuli were calculated and applied simultaneously.

Results

Inhibitory interneuron plasticity can mediate learned irrelevance

The first simulations established that the model could learn to distinguish relevant versus irrelevant stimuli based on whether or not the stimuli were reinforced. The network was presented with two independent, 200 ms long stimuli with inter-trial-intervals that were independently jittered between 20-30 seconds (**Figure 2A**, *colored boxes on bottom*). The conditioned stimulus (CS+) was paired with an unconditioned stimulus (US) that began 100 ms after CS onset and terminated 100 ms following offset; the second stimulus was not paired with a US (CS-). Initially, the increased input from the Sensory

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

layer, $\mathbf{x}(t)$, associated with presentation of either stimulus evoked an increased response among Cortex excitatory units ($\mathbf{E}(t)$) and, by extension, the population-based salience signal ($S(t)$; **Figure 2A**, *left*). As the number of presentations accumulated, there was a selective reduction in the degree to which excitatory units responded to the CS-, but no reduction in the degree to which the network responded to the CS+ (**Figure 2A**, *right*). **Figure 2B** illustrates the gradual decrease in excitatory unit population responses to the CS- (*left*) and the corresponding increases in the Cortex inhibitory unit ($\mathbf{I}(t)$) response (*right*).

The increased responsiveness of the Cortex inhibitory unit to the CS- over presentations was due to the gradually increased connection weights between the units of the Sensory layer and the Cortex-inhibitory unit ($W^{\mathbf{x}\rightarrow\mathbf{I}}$), caused by the temporal difference learning rule. We next examined whether the same patterns could be observed using other model versions, in which synapses either between Sensory and Cortex-excitatory units ($W^{\mathbf{x}\rightarrow\mathbf{E}}$), or between Cortex-inhibitory and excitatory units ($W^{\mathbf{I}\rightarrow\mathbf{E}}$) were modified. This comparison included an examination of how the model responded to disrupted inhibition (see Methods). The results of these tests are described in **Figure 2C**. Both $W^{\mathbf{x}\rightarrow\mathbf{I}}$ plasticity and $W^{\mathbf{x}\rightarrow\mathbf{E}}$ plasticity models exhibited much greater ability to learn to distinguish relevant versus irrelevant stimuli, indicated by the salience signal ($S(t)$) during the CS+ relative to CS-. Both of these models were better at distinguishing CS+ and CS- than the $W^{\mathbf{I}\rightarrow\mathbf{E}}$ plasticity model. However, disrupted inhibition only eliminated learned irrelevance in the $W^{\mathbf{x}\rightarrow\mathbf{I}}$ plasticity model. Differences in how the model types responded to disrupted inhibition could be assessed statistically: even ten repetitions of the simulation was more than sufficient to demonstrate an interaction effect between model type and inhibitory manipulation (two-way ANOVA, type $\times$ manipulation: $F_{(2,54)} = 21.85, p = 10^{-}07$; one-way ANOVA comparing the disrupted inhibition conditions: $F_{(2,27)} = 21.65, p = 2x10^{-}06$, multiple comparisons between all groups significantly different using a Bonferroni correction).

If we take as an assumption that learned irrelevance impairments found in schizophrenia or other disorders arise from changes in inhibitory neurons, consistent with observed pharmacological effects in rodents, then the patterns observed in the $W^{\mathbf{x}\rightarrow\mathbf{I}}$ plasticity model offer the best fit to the pathology. This supports the hypothesis that relevance learning is mediated by inhibitory interneuron plasticity.

Inhibitory interneuron plasticity can explain blocking data

Another well-established relevance learning phenomenon is 'blocking', in which one stimulus that has been previously reinforced can occlude learning for another reinforced stimulus Kamin, L J. Blocking is also known to be affected in schizophrenia (Jones et al., 1992, 1997; Oades et al., 2000). To examine whether the model exhibited blocking, a standard blocking protocol was simulated, illustrated in **Figure 3A**. Two different stimuli were presented to the network, CS-A (non-blocked) and CS-B (blocked). The difference between the non-blocked, CS-A, stimulus and the blocked, CS-B, stimulus is that CS-A was conditioned alone with the US (following habituation pre-exposures) while CS-B was conditioned only when paired with CS-A (following pre-exposures and CS-A conditioning). When this type of protocol is used in either rodent (e.g., Kamin, L J; Ganesan and Pearce) or human (reviewed by De Houwer et al.) experiments, it leads to CS-A being recognized as relevant for reward/punishment, but CS-B being judged irrelevant. The blocking effect was measured in the model by comparing the response of the excitatory unit population to CS-A and CS-B during the final test sessions (**Figure 3B**, also **Figure 3C** *inset*). As predicted, the model exhibited the basic blocking effect seen in people and animals, with CS-A generating a large excitatory response and CS-B generating a small one (**Figure 3C**, *left inset*). Because learning was supported by inhibitory neuron plasticity (in this, the $W^{\mathbf{x}\rightarrow\mathbf{I}}$ plasticity model), changes in population responses to both CS-A and CS-B over stimulus presentations paralleled changes in inhibitory neuron, $\mathbf{I}(\mathbf{t})$ responses (**Figure 3C**, *right inset*). Notably, a strong increase in inhibitory neuron activity was observed during the "blocking" phase of conditioning, reflecting feed-forward inhibition compensating for both stimuli being presented simultaneously (**Figure 3C**, *presentations 0-50*). We note that this is also consistent with observations of increased fast-spiking, narrow-waveform neuron activity during stimulus presentations and movement (Insel and Barnes, 2015).

As with the learned irrelevance simulation, blocking and the effect of inhibitory disruptions was

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

tested in different versions of the models, defined by which synapses ($W^{\mathbf{x}\to\mathbf{I}}$, $W^{\mathbf{x}\to\mathbf{E}}$, $W^{\mathbf{E}\to\mathbf{I}}$) were plastic. Consistent with predictions, the blocking effect was eliminated in the $W^{\mathbf{x}\to\mathbf{I}}$ plasticity model, following even a 10% disruption of inhibition (**Figure 3D**, *left bars*). Blocking was also observed in the $W^{\mathbf{x}\to\mathbf{E}}$ model; however, in this model version an unexpected "reverse blocking" effect was observed following inhibitory disruptions (**Figure 3D**, *middle bars*). This was likely due to the supervised-learning mechanism becoming over-active following inhibitory disruptions, leading to a reduction in synaptic weights corresponding with CS-A presentations (regardless of it being paired with the US). No blocking effects could be obtained in the $W^{\mathbf{I}\to\mathbf{E}}$ model (**Figure 3D**, *right bars*). Once again, these results could be judged statistically, with 10 repetitions more than sufficient to reveal an interaction between model type and manipulation (two-way ANOVA, $F_{(2,54)} = 21.85, p = 10^{-}07$), one-way ANOVA of disrupted inhibition condition: $F_{(2,27)} = 21.65, p = 2x10^{-}06$, with multiple comparisons test with Bonferroni correction showing a significant difference between $W^{\mathbf{x}\to\mathbf{I}}$ and $W^{\mathbf{x}\to\mathbf{E}}$ models).

Consistent with the results from the learned irrelevance simulations, the pattern of effects in the blocking paradigm, when considered in the context of schizophrenia symptoms and inhibitory neuron pathologies, lend credence to the hypothesis that blocking is mediated by plasticity in inhibitory neurons.

Inhibitory relevance-learning network with amygdala module recapitulates effect of pharmacological manipulations on latent inhibition

Motivated in part by the pathology of schizophrenia, researchers have previously sought to experimentally link EI balance in frontal cortex to various forms of conditioning (Piantadosi and Floresco, 2014; Courtin et al., 2014). A set of simulations was designed to determine whether our model could recapitulate the findings of these studies. To explicitly simulate the link between experimental manipulations of the mPFC and behavior, it was necessary to demonstrate how the output of the mPFC, and in particular the relevance signal, $S(t)$, might affect an efferent region that directly controls behavioral output. Since many of the rodent studies in this area use fear conditioning, the output layer was designed to represent the mamalian amygdala, and the levels of 'Amygdala' unit activity were equated with fear expression. The Amygdala layer was modeled as a competitive network with 10 units receiving input directly from Cortex excitatory units (see Methods).

The first simulations examined the findings of Piantadosi and Floresco (2014), which showed that a GABA-A receptor antagonist, applied to the medial prefrontal cortex (mPFC), can have different effects on latent inhibition when applied at different phases of the learning protocol. As described above, latent inhibition refers to the phenomenon wherein it is harder to associate a stimulus with a reinforcer if a subject has previously been exposed to that stimulus. In the study by Piantadosi and Floresco (2014), animals were separated into two groups: those that received pre-exposures of the CS (PE) and those that did not (NPE). When the CS was subsequently paired with a footshock, the PE group was less likely to learn the fear association compared with the NPE group (i.e. the animals exhibited latent inhibition). Importantly, the authors found that blocking GABA-A receptors had different effects if done during the conditioning period or during the test: GABA-A antagonists infused during conditioning amplified latent inhibition, whereas infusions during testing disrupted latent inhibition (**Figure 4A**). These experiments were simulated using a 20% reduction in inhibition (as per the learned irrelevance and blocking simulations) to mimic blockade of GABA-A receptors (see Methods). The model showed a similar pattern of results as observed by Piantadosi and Floresco (2014), with simulated GABA-A blockade increasing the latent inhibition effect if applied during conditioning, and eliminating latent inhibition if applied during testing (**Figure 4B**).

The link between the salience signal (i.e., levels of excitatory unit activity) and Amygdala activity can be better understood by examining how each changed from one trial to the next (**Figures 4C-D**). As illustrated by the downward slope of activity among Cortex excitatory units over pre-exposure trials, the network receiving pre-exposures learned to ignore the CS over the preconditioning period, so that during the conditioning period (gray shaded area in **Figures 4C** and **4D**) Cortex activity was too low to always push the Amygdala past threshold, θ, making it less likely to trigger learning in the Amygdala competitive network (see equations (10)). The salience signal during conditioning for the

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

non- pre-exposure tests of the model, however, resulted in a sufficiently strong Amygdala response to the CS to induce learning in the competitive network. When inhibitory signaling was experimentally reduced during conditioning, both Cortex activity and Amygdala learning were amplified. However, this learning included many excitatory units that were not coding the CS, so that the PE network remained relatively inactive during testing (**Figures 4C-D**). In contrast, when inhibitory signaling was reduced during testing, many Cortex units become active in both PE and NPE conditions. Since the Amygdala was modeled as a competitive network, the overactivity of non-CS coding inputs interfered with the activity of CS-coding inputs, resulting in overall low Amygdala responses in both PE and NPE conditions (**Figures 4C-D**). These results provide a new interpretation of the Piantadosi and Floresco (2014) data, one that that emphasizes the importance of the interaction between a relevance code in the mPFC and learning rates in a second system such as the amygdala.

Inhibitory relevance-learning network with amygdala module recapitulates effect of optogenetic manipulations on fear behavior

The second simulation of experimental results addressed work by Courtin et al. (2014), which examined how the activity of PV+ interneurons in the mPFC controls fear expression. As described in Methods, PV+ interneurons are the cells modeled by the Cortex inhibitory unit, $\mathbf{I}(t)$. We simulated the experiments using the same network and parameters, including the Amygdala layer, as used to simulate latent inhibition above (see Methods). The original study by Courtin et al. (2014) demonstrated that optogenetic stimulation of PV+ interneurons in the mPFC resulted in increased freezing activity in mice, both before conditioning and, even more prominently, when stimulation was paired with a CS following extinction (**Figure 5A**, *left*). When we applied the same protocol on the model, using a reduction in inhibitory inputs to mimic optogenetic silencing (see Methods), the same pattern of activity was observed in the Amygdala layer (**Figure 5B**, *left*). The original experiments also showed that activation of mPFC PV+ interneurons decreased freezing to a conditioned CS (**Figure 5A**, *right*). This was consistent with activity patterns in the model in a subsequent set of simulations (**Figure 5B**, *right*). These results further support the hypothesis that the effects of the mPFC on efferent regions that control behavior is related to a potential relevance code mediated by feed-forward, divisive inhibition. They also offer evidence that the present model, in spite of its simplicity, captures an essential relationship between the role of inhibition in the mPFC region and the competitive network in the amygdala (Rashid et al., 2016).

Relevance learning can be multiplexed with input classification

A final set of simulations was used to investigate a key computational advantage to using inhibitory neuron plasticity for relevance learning, specifically by providing a method by which the brain could multiplex relevance learning with learning about other stimulus information. If the overall level of excitatory activity encodes relevance (via $S(t)$), and this is controlled by feedforward inhibition, then the excitatory synapses in the network should still be free to control the specific pattern of $\mathbf{E}(t)$ to encode other information. This can be described mathematically by viewing the excitatory Cortex activity patterns $\mathbf{E}(t)$ as vectors, where the norm (length) of the vector is a signal of relevance, but the position that the vector points in encodes other aspects of a stimulus, such as orientation, frequency, or category.

To test this idea, the network was trained to categorize 10 different stimulus classes, with only one of these paired with a reward. The prediction was that the network could learn information about relevance and also learn to respond with output patterns specific to the correct stimulus category. To train the network to categorize stimuli, we employed a softmax output layer (see Methods) and trained the excitatory pathway in the network with backpropagation of error (Rumelhart et al., 1986) (**Figure 6A**). It is worth noting that although backpropagation of error is traditionally not seen as a biologically realistic learning algorithm, there is evidence that it could be approximated with biologically realistic mechanisms (Scellier and Bengio, 2017; Lillicrap et al., 2016; Guergiuev et al., 2016; Zenke and Ganguli, 2017). Furthermore, independent of the specific algorithm used, the goal of the simulation was to offer a proof of the multiplexing concept. Training of the excitatory pathway with backpropagation was done

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

concurrently with training of the feedforward inhibition pathway using the relevance learning algorithm (as described in Methods). Over the course of training, the network learned to dissociate the rewarded stimulus category from the unrewarded ones, via the relevance signal, $S(t)$ (**Figure 6B**). Importantly, $S(t)$ did not differentiate between the unrewarded categories (**Figure 6B**, *red lines*), demonstrating that it was not encoding the categories, *per se*, but their relevance for predicting reward. At the same time, the set of 'Category' output units did learn to differentiate all 10 categories of stimuli. A cross-entropy loss function was used to evaluate the success of categorization, with lower values indicating a higher degree of separation between the categories. Over the course of 20 simulated seconds of training, this measure dropped to almost zero and the output layer was achieving roughly 95% accuracy on average (**Figure 6C**). Importantly, relevance learning and category learning were operating simultaneously. These results demonstrate the potential for multiplexing relevance signals with other stimulus information by using the overall level of excitatory activity as a code for value.

Discussion

The simulations presented here explored the link between relevance assignment and neural disinhibition. They were motivated, in part, by the hypothesis that relevance learning impairments, much like those observed in schizophrenia, can be caused by a disruption to feed-forward inhibitory processes.The model was structured as simply as was necessary to examine this hypothesis, incorporating three core mechanisms. First, relevance was coded as the overall excitatory activity in the 'Cortex' layer. Specifically, the norm (length) of the excitatory units' activity vector was treated as a reinforcement learning value function, though "unsigned" in that it treated positive (reward) and negative (punishment) values equivalently (Figure 1). Second, the model used feed-forward inhibition–i.e., the connections from the 'Sensory' input layer to Cortex inhibitory units–to control the overall level of Cortex excitatory activity. When paired with the first mechanism, the result was that disruptions to inhibition led to failures in relevance attribution (Figures 2 & 3). Third, the model used an established reinforcement learning algorithm, temporal difference learning, to train the feed-forward inhibitory connections and thereby learn to differentiate relevant versus irrelevant stimuli. When these mechanisms were further connected in sequence with a an output (the 'Amygdala') that used a competitive learning mechanism, this led to specific predictions about how disruptions to inhibition would alter fear behavior (Figure 4).

These three mechanisms are highly consistent with previous empirical work. The first mechanism, linking overall levels of excitation with a code for unsigned value, matches recent data on the importance of mPFC disinhibition for coding relevant situations (Pi et al., 2013; Courtin et al., 2014; Pinto and Dan, 2015; Kim et al., 2016a,b), including the observation that net levels of activity in putative pyramidal neurons increases near reward sites (Insel and Barnes, 2015). The second mechanism, assigning control of this relevance code to feed-forward inhibition, matches empirical findings on the importance of inhibition for behaviors like latent inhibition (e.g. (Piantadosi and Floresco, 2014)). It also matches decades of work linking relevance impairments in schizophrenia (Mcghie and Chapman, 1961; Venables, 1964; Lang and Buss, 1965; McGhie, 1970; Garmezy, 1977; Johnson, 1985; Kapur, 2003) with evidence that inhibitory neurons, and in particular classes of inhibitory neurons supporting feed-forward inhibition, may be differentially compromised in the disease (Benes and Berretta, 2001; Lewis et al., 2005; Lewis and Moghaddam, 2006; Lewis, 2014; Gonzalez-Burgos et al., 2015; Krystal et al., 2017). Finally, the third mechanism, wherein inhibitory interneuron plasticity is the means for learning to differentiate relevant versus irrelevant stimuli, is consistent with findings that the neural connections supporting feed-forward inhibition are plastic (Kullmann et al., 2012; Kullmann and Lamsa, 2011; Bartos et al., 2011), in some cases requiring NMDA receptors with a well established importance for associative plasticity (Lamsa et al., 2007, 2005; Le Roux et al., 2013). There has been considerable recent interest in understanding the functions of inhibitory neuron plasticity (Barron et al.; Sprekeler; Hennequin et al.). These three mechanisms together comprise a more general theory of cortical coding: inhibitory plasticity may act in parallel with excitatory plasticity to accomplish different learning functions–while excitatory plasticity could provide a mechanism for learning various tasks, such as categorization, inhibitory plasticity could

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

simultaneously provide a mechanism for relevance learning (Figure 5). This suggests a multiplexed code in the neocortex, and brings together the known specificity of inhibitory neuron plasticity with theories about their importance for maintaining EI balance.

As this was an abstract neural network model, many features of the real brain were absent. The most notable was the absence of feedback connections within Cortex. By excluding these connections, the mathematical complexity of calculating synaptic balances and their experience-dependent changes could be minimized, and it became possible to isolate those processes sufficiently to develop the learning algorithms that could explain the behavioral phenomena of interest. The results demonstrated that plasticity in the synapses connecting inputs to inhibitory neurons is sufficient to support relevance learning. Such a mechanism also causes relevance learning to be dysfunctional following disrupted inhibition. Other authors have taken a more biologically realistic approach to addressing the function of inhibitory neurons for learning and memory, including adding organized, feedback excitatory connections (Murray et al., 2014), along with organized, "motif" connections of specific inhibitory neuron types (Yang et al., 2016). The temporal precision of feed-forward inhibitory connections has also been proposed to regulate plasticity in excitatory neurons (). The findings from these more complex networks are consistent with those presently observed, and it seems likely that there is a meaningful mapping between the two. The present emphasis on simplicity for the sake of mathematical tractability offers only a first step toward understanding and treating disorders like schizophrenia.

In addition to excluding certain properties of the real brain, some other features of the model were kept abstract in order to remain agnostic as to their potential biological implementation. For example, the computed salience signal ($S(t)$) and corresponding prediction error responsible for maintaining EI balance ($\beta(t)$) might be integrated locally within the cortex itself (e.g. by feedback connections from excitatory to local inhibitory neurons), or computed by subcortical regions. As described in the Introduction, it is well known that networks maintain a strong balance between excitation and inhibition (Wehr and Zador, 2003; Zhang et al., 2003; Dorrn et al., 2010; Gabernet et al., 2005; Wilent and Contreras, 2005; Daw et al., 2007; Chittajallu and Isaac, 2010; Poo and Isaacson, 2009; Anderson et al., 2000; Xue et al., 2014; Haider et al., 2006) (see also the review by (Froemke, 2015)). While the process by which networks maintain this balance is not fully known, it has been proposed to be supported by plasticity at inhibitory synapses in response to feedback excitatory signals (Vogels et al., 2011; Luz and Shamir, 2012). Recent work has demonstrated local changes in synaptic scaling at both inhibitory and excitatory synapses following local changes in excitation (Barral and D Reyes, 2016). One possibility is that while the local network is capable of maintaining EI balance, the EI balance set-point can be adjusted by signals from extrinsic neuromodulators. Many neuromodulators are capable of signaling reinforcement or salience. While dopamine has received much of the attention in this regard, acetylcholine has also recently come under the spotlight (e.g. (Kuchibhotla et al., 2017; Letzkus et al., 2015, 2011; Hangya et al., 2015)). Disinhibition may also involve a particular class of inhibitory neurons that contain vasoactive intestinal polypeptide (VIP+ interneurons (Pi et al., 2013)). Although the present model does not identify the specific biological process responsible for transmitting the salience and prediction error signals to inhibitory neuron synapses, it does predict that the plasticity processes responsible for relevance learning directly impact the mechanisms for maintaining EI balance; otherwise, EI balance maintenance would be constantly working to compensate for the adjustments to EI balance associated with relevance learning. To accurately capture the mechanisms of the supervisory process, it will likely be necessary to increase the complexity of the model by including feedback connections between excitatory neurons, and connections from excitatory to inhibitory neurons.

The primary prediction of the present work is that relevance learning involves plasticity on the feed-forward, input-to-inhibitory neuron synapses in a cortical circuit. This prediction is consistent with recent data showing that the number of excitatory synapses onto parvalbumin-expressing inhibitory neurons is reduced in schizophrenia (Chung et al., 2016). It also contributes to an idea gaining momentum that during memory storage patterns of inhibitory modifications mirror excitatory modifications, in order to maintain EI balance and reduce inappropriate recall, i.e. that there is an "inhibitory engram" (Barron et al.). The present model has many parallels with this idea. Specifically, inhibitory synapses in the model learned to match any increases in excitatory synapses in order to keep the 'Cortex' output at

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the homeostatic set-point. Only when the prediction error signal indicated that an input was relevant was this mirrored inhibition relaxed to allow the excitatory activity to increase.

The simulations also addressed how the proposed code for relevance might impact learning and activity in a post-synaptic region. With the aim of simulating experimental fear-learning data, the post-synaptic region chosen was the amygdala, which was modeled as a competitive learning network (Rumelhart and Zipser, 1985). Use of a competitive network is consistent with known properties of the mammalian amygdala (Yiu et al., 2014; Rashid et al., 2016). By combining our proposed relevance code at one layer of the model with a competitive learning rule at the next, the model was capable of responding selectively to only those specific stimuli that had been paired with reinforcement signals in the past. The model thereby became capable of replicating behavioral patterns from both pharmacological (Piantadosi and Floresco, 2014) and optogenetic (Courtin et al., 2014) manipulations to mPFC inhibitory neurons.

One area that was not explored in the current set of experiments was the increasingly apparent link between pathologies of EI balance and deficits in social behavior and motivation (Rubenstein and Merzenich, 2003; Yizhar et al., 2011; Nelson and Valakh, 2015). It seems likely, however, that some of the more basic results from the present investigations could offer more understanding for why some individuals have more difficulty filtering or dynamically processing social information. Tackling these complex problems will require a convergence of multiple experimental and theoretical approaches, and mathematically tractable network models that include excitatory-inhibitory interactions will be an essential tool.

Altogether, our theoretical investigations provide a potential explanation for why behaviors such as gating and relevance learning may emerge from disrupted inhibition, and therefore, how pathologies of inhibition may underlie neuropsychiatric conditions such as schizophrenia. The ideas are not entirely new, in many ways they reformulate a long existing hypothesis that schizophrenia is a disruption of feed-forward inhibition (Johnson, 1985). But the model offers a computational description of the process with defined links between several functional elements. Furthermore, it offers valuable predictions about the importance of plasticity in both excitatory and inhibitory neurons, lending insights into the normally functioning brain.

Acknowledgements

We would like to thank Adam Santoro for his comments on an earlier draft of this manuscript. This work was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant RGPIN-04947), Google (Faculty Research Award 2016), and the Canadian Institute for Advances Research (Learning in Machines and Brains Fellowship).

References

Agmon, A. and Connors, B. W. (1992). Correlation between intrinsic firing patterns and thalamocortical synaptic responses of neurons in mouse barrel cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 12(1):319–329.

Anderson, J. S., Carandini, M., and Ferster, D. (2000). Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *Journal of Neurophysiology*, 84(2):909–926.

Barral, J. and D Reyes, A. (2016). Synaptic scaling rule preserves excitatory-inhibitory balance and salient neuronal network dynamics. *Nature Neuroscience*, 19(12):1690–1696.

Barron, H. C., Vogels, T. P., Behrens, T. E., and Ramaswami, M. Inhibitory engrams in perception and memory. 114(26):6666–6674.

Bartos, M., Alle, H., and Vida, I. (2011). Role of microcircuit structure and input integration in hippocampal interneuron recruitment and plasticity. *Neuropharmacology*, 60(5):730–739.

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Baruch, I., Hemsley, D. R., and Gray, J. A. (1988). Differential performance of acute and chronic schizophrenics in a latent inhibition task. *The Journal of Nervous and Mental Disease*, 176(10):598–606.

Benes, F. M. and Berretta, S. (2001). GABAergic interneurons: implications for understanding schizophrenia and bipolar disorder. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 25(1):1–27.

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, 68(5):815–834.

Carcea, I. and Froemke, R. C. (2013). Cortical plasticity, excitatory-inhibitory balance, and sensory perception. *Progress in Brain Research*, 207:65–90.

Chittajallu, R. and Isaac, J. T. R. (2010). Emergence of cortical inhibition by coordinated sensory-driven plasticity at distinct synaptic loci. *Nature Neuroscience*, 13(10):1240–1248.

Chung, D. W., Fish, K. N., and Lewis, D. A. (2016). Pathological basis for deficient excitatory drive to cortical parvalbumin interneurons in schizophrenia. *The American Journal of Psychiatry*, 173(11):1131–1139.

Courtin, J., Chaudun, F., Rozeske, R. R., Karalis, N., Gonzalez-Campo, C., Wurtz, H., Abdi, A., Baufreton, J., Bienvenu, T. C. M., and Herry, C. (2014). Prefrontal parvalbumin interneurons shape neuronal activity to drive fear expression. *Nature*, 505(7481):92–96.

Dantzker, J. L. and Callaway, E. M. (2000). Laminar sources of synaptic input to cortical inhibitory interneurons and pyramidal neurons. *Nature Neuroscience*, 3(7):701–707.

Daw, M. I., Scott, H. L., and Isaac, J. T. R. (2007). Developmental synaptic plasticity at the thalamocortical input to barrel cortex: mechanisms and roles. *Molecular and Cellular Neurosciences*, 34(4):493–502.

De Houwer, J., Thomas, S., and Baeyens, F. Associative learning of likes and dislikes: a review of 25 years of research on human evaluative conditioning. 127(6):853–869.

Dorrn, A. L., Yuan, K., Barker, A. J., Schreiner, C. E., and Froemke, R. C. (2010). Developmental sensory experience balances cortical excitation and inhibition. *Nature*, 465(7300):932–936.

Enomoto, T., Tse, M. T., and Floresco, S. B. (2011). Reducing prefrontal gamma-aminobutyric acid activity induces cognitive, behavioral, and dopaminergic abnormalities that resemble schizophrenia. *Biological Psychiatry*, 69(5):432–441.

Fino, E., Packer, A. M., and Yuste, R. (2013). The Logic of Inhibitory Connectivity in the Neocortex. *The Neuroscientist*, 19(3):228–237.

Froemke, R. C. (2015). Plasticity of cortical excitatory-inhibitory balance. *Annual Review of Neuroscience*, 38:195–219.

Gabernet, L., Jadhav, S. P., Feldman, D. E., Carandini, M., and Scanziani, M. (2005). Somatosensory integration controlled by dynamic thalamocortical feed-forward inhibition. *Neuron*, 48(2):315–327.

Ganesan, R. and Pearce, J. M. Effect of changing the unconditioned stimulus on appetitive blocking. 14(3):280–291.

Garmezy, N. (1977). The psychology and psychopathology of attention. *Schizophrenia Bulletin*, 3(3):360–369.

Gibson, J. R., Beierlein, M., and Connors, B. W. (1999). Two networks of electrically coupled inhibitory neurons in neocortex. *Nature*, 402(6757):75–79.

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Gonchar, Y. and Burkhalter, A. (2003). Distinct GABAergic targets of feedforward and feedback connections between lower and higher areas of rat visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 23(34):10904–10912.

Gonzalez-Burgos, G., Cho, R. Y., and Lewis, D. A. (2015). Alterations in cortical network oscillations and parvalbumin neurons in schizophrenia. *Biological Psychiatry*, 77(12):1031–1040.

Gray, N. S. and Snowden, R. J. (2005). The relevance of irrelevance to schizophrenia. *Neuroscience and Biobehavioral Reviews*, 29(6):989–999.

Guergiuev, J., Lillicrap, T. P., and Richards, B. A. (2016). Deep learning with segregated dendrites. *arXiv preprint arXiv:1610.00161*.

Haider, B., Duque, A., Hasenstaub, A. R., and McCormick, D. A. (2006). Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 26(17):4535–4545.

Han, J.-H., Kushner, S. A., Yiu, A. P., Cole, C. J., Matynia, A., Brown, R. A., Neve, R. L., Guzowski, J. F., Silva, A. J., and Josselyn, S. A. (2007). Neuronal competition and selection during memory formation. *Science*, 316(5823):457–460.

Hangya, B., Ranade, S., Lorenc, M., and Kepecs, A. (2015). Central Cholinergic Neurons Are Rapidly Recruited by Reinforcement Feedback. *Cell*, 162(5):1155–1168.

Hennequin, G., Agnes, E. J., and Vogels, T. P. Inhibitory plasticity: Balance, control, and codependence.

Insel, N. and Barnes, C. A. (2015). Differential activation of fast-spiking and regular-firing neuron populations during movement and reward in the dorsal medial frontal cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, 25(9):2631–2647.

Isaacson, J. S. and Scanziani, M. (2011). How inhibition shapes cortical activity. *Neuron*, 72(2):231–243.

Johnson, J. D. (1985). A mechanism to inhibit input activation and its dysfunction in schizophrenia. *The British Journal of Psychiatry: The Journal of Mental Science*, 146:429–435.

Jones, S. H., Gray, J. A., and Hemsley, D. R. (1992). Loss of the kamin blocking effect in acute but not chronic schizophrenics. *Biological Psychiatry*, 32(9):739–755.

Jones, S. H., Hemsley, D., Ball, S., and Serra, A. (1997). Disruption of the kamin blocking effect in schizophrenia and in normal subjects following amphetamine. *Behavioural Brain Research*, 88(1):103–114.

Kamin, L J. Predictability, surprise, attention, and conditioning. In *Punishment and aversive behavior*, pages 242–259. Appleton-Century-Crofts.

Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *The American Journal of Psychiatry*, 160(1):13–23.

Karnani, M. M., Agetsuma, M., and Yuste, R. (2014). A blanket of inhibition: functional inferences from dense inhibitory connectivity. *SI: Inhibition: Synapses, Neurons and Circuits*, 26(0):96–102.

Kato, H. K., Gillet, S. N., and Isaacson, J. S. (2015). Flexible sensory representations in auditory cortex driven by behavioral relevance. *Neuron*, 88(5):1027–1039.

Kim, D., Jeong, H., Lee, J., Ghim, J.-W., Her, E. S., Lee, S.-H., and Jung, M. W. (2016a). Distinct roles of parvalbumin- and somatostatin-expressing interneurons in working memory. *Neuron*, 92(4):902–915.

Kim, H., Ährlund Richter, S., Wang, X., Deisseroth, K., and Carlén, M. (2016b). Prefrontal parvalbumin neurons in control of attention. *Cell*, 164(1):208–218.

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Krystal, J. H., Anticevic, A., Yang, G. J., Dragoi, G., Driesen, N. R., Wang, X.-J., and Murray, J. D. (2017). Impaired tuning of neural ensembles and the pathophysiology of schizophrenia: A translational and computational neuroscience perspective. *Biological Psychiatry*, 81(10):874–885.

Kuchibhotla, K. V., Gill, J. V., Lindsay, G. W., Papadoyannis, E. S., Field, R. E., Sten, T. A. H., Miller, K. D., and Froemke, R. C. (2017). Parallel processing by cortical inhibition enables context-dependent behavior. *Nature Neuroscience*, 20(1):62–71.

Kullmann, D., Moreau, A., Bakiri, Y., and Nicholson, E. (2012). Plasticity of Inhibition. *Neuron*, 75(6):951–962.

Kullmann, D. M. and Lamsa, K. P. (2011). LTP and LTD in cortical GABAergic interneurons: Emerging rules and roles. *Neuropharmacology*, 60(5):712–719.

Lamsa, K., Heeroma, J. H., and Kullmann, D. M. (2005). Hebbian LTP in feed-forward inhibitory interneurons and the temporal fidelity of input discrimination. *Nature Neuroscience*, 8(7):916–924.

Lamsa, K., Irvine, E. E., Giese, K. P., and Kullmann, D. M. (2007). NMDA receptor-dependent long-term potentiation in mouse hippocampal interneurons shows a unique dependence on Ca^{2+} /calmodulin-dependent kinases: NMDA receptor-dependent LTP in interneurons is independent of αCAMKII. *The Journal of Physiology*, 584(3):885–894.

Lang, P. J. and Buss, A. H. (1965). PSYCHOLOGICAL DEFICIT IN SCHIZOPHRENIA. II. INTERFERENCE AND ACTIVATION. *Journal of Abnormal Psychology*, 70:77–106.

Le Roux, N., Cabezas, C., B?hm, U. L., and Poncer, J. C. (2013). Input-specific learning rules at excitatory synapses onto hippocampal parvalbumin-expressing interneurons: Long term plasticity in parvalbumin interneurons. *The Journal of Physiology*, 591(7):1809–1822.

Letzkus, J. J., Wolff, S. B. E., and Lüthi, A. (2015). Disinhibition, a circuit mechanism for associative learning and memory. *Neuron*, 88(2):264–276.

Letzkus, J. J., Wolff, S. B. E., Meyer, E. M. M., Tovote, P., Courtin, J., Herry, C., and Lüthi, A. (2011). A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature*, 480(7377):331–335.

Lewis, D. A. (2014). Inhibitory neurons in human cortical circuits: substrate for cognitive dysfunction in schizophrenia. *Current Opinion in Neurobiology*, 26:22–26.

Lewis, D. A., Hashimoto, T., and Volk, D. W. (2005). Cortical inhibitory neurons and schizophrenia. *Nature Reviews. Neuroscience*, 6(4):312–324.

Lewis, D. A. and Moghaddam, B. (2006). Cognitive dysfunction in schizophrenia: convergence of gamma-aminobutyric acid and glutamate alterations. *Archives of Neurology*, 63(10):1372–1376.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7:13276.

Lubow, R. E. and Moore, A. U. (1959). Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. *Journal of Comparative and Physiological Psychology*, 52:415–419.

Lubow, R. E., Weiner, I., Schlossberg, A., and Baruch, I. (1987-06). Latent inhibition and schizophrenia. *Bulletin of the Psychonomic Society*, 25(6):464–467.

Luz, Y. and Shamir, M. (2012). Balancing feed-forward excitation and inhibition via hebbian inhibitory synaptic plasticity. *PLoS computational biology*, 8(1):e1002334.

McAllister, K. H. (1997). A single administration of d-amphetamine prior to stimulus pre-exposure and conditioning attenuates latent inhibition. *Psychopharmacology*, 130(2):79–84.

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

McGhie, A. (1970). Attention and perception in schizophrenia. *Progress in Experimental Personality Research*, 5:1–35.

Mcghie, A. and Chapman, J. (1961). Disorders of attention and perception in early schizophrenia. *The British Journal of Medical Psychology*, 34:103–116.

Murray, J. D., Anticevic, A., Gancsos, M., Ichinose, M., Corlett, P. R., Krystal, J. H., and Wang, X.-J. (2014). Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cerebral Cortex (New York, N.Y.: 1991)*, 24(4):859–872.

Nelson, S. B. and Valakh, V. (2015). Excitatory/inhibitory balance and circuit homeostasis in autism spectrum disorders. *Neuron*, 87(4):684–698.

Oades, R. D., Rao, M. L., Bender, S., Sartory, G., and Müller, B. W. (2000). Neuropsychological and conditioned blocking performance in patients with schizophrenia: assessment of the contribution of neuroleptic dose, serum levels and dopamine d2-receptor occupancy. *Behavioural Pharmacology*, 11(3):317–330.

Paninski, L., Pillow, J., and Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. In Paul Cisek, T. D. a. J. F. K., editor, *Progress in Brain Research*, volume Volume 165, pages 493–507. Elsevier.

Pi, H.-J., Hangya, B., Kvitsiani, D., Sanders, J. I., Huang, Z. J., and Kepecs, A. (2013). Cortical interneurons that specialize in disinhibitory control. *Nature*, 503(7477):521–524.

Piantadosi, P. T. and Floresco, S. B. (2014). Prefrontal cortical GABA transmission modulates discrimination and latent inhibition of conditioned fear: relevance for schizophrenia. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 39(10):2473–2484.

Pinto, L. and Dan, Y. (2015). Cell-type-specific activity in prefrontal cortex during goal-directed behavior. *Neuron*, 87(2):437–450.

Poo, C. and Isaacson, J. S. (2009). Odor representations in olfactory cortex: "sparse" coding, global inhibition, and oscillations. *Neuron*, 62(6):850–861.

Rascle, C., Mazas, O., Vaiva, G., Tournant, M., Raybois, O., Goudemand, M., and Thomas, P. (2001). Clinical features of latent inhibition in schizophrenia. *Schizophrenia Research*, 51(2):149–161.

Rashid, A. J., Yan, C., Mercaldo, V., Hsiang, H.-L. L., Park, S., Cole, C. J., De Cristofaro, A., Yu, J., Ramakrishnan, C., Lee, S. Y., Deisseroth, K., Frankland, P. W., and Josselyn, S. A. (2016). Competition between engrams influences fear memory formation and recall. *Science (New York, N.Y.)*, 353(6297):383–387.

Rubenstein, J. L. R. and Merzenich, M. M. (2003). Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes, Brain, and Behavior*, 2(5):255–267.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(9):533–536.

Rumelhart, D. E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, 9(1):75–112.

Scellier, B. and Bengio, Y. (2017). Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Frontiers in Computational Neuroscience*, 11:24.

Shagass, C. (1976). An electrophysiological view of schizophrenia. *Biological Psychiatry*, 11(1):3–30.

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Sprekeler, H. Functional consequences of inhibitory plasticity: homeostasis, the excitation-inhibition balance and beyond. 43:198–203.

Sutton, R. S. and Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT Press.

Venables, P. H. (1964). INPUT DYSFUNCTION IN SCHIZOPHRENIA. *Progress in Experimental Personality Research*, 72:1–47.

Venables, P. H. (1977). The electrodermal psychophysiology of schizophrenics and children at risk for schizophrenia: controversies and developments. *Schizophrenia Bulletin*, 3(1):28–48.

Vogels, T. P. and Abbott, L. F. (2007). Gating deficits in model networks: a path to schizophrenia? *Pharmacopsychiatry*, 40 Suppl 1:S73–77.

Vogels, T. P. and Abbott, L. F. (2009). Gating multiple signals through detailed balance of excitation and inhibition in spiking networks. *Nature Neuroscience*, 12(4):483–491.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science (New York, N.Y.)*, 334(6062):1569–1573.

Wehr, M. and Zador, A. M. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature*, 426(6965):442–446.

Wilent, W. B. and Contreras, D. (2005). Dynamics of excitation and inhibition underlying stimulus selectivity in rat somatosensory cortex. *Nature Neuroscience*, 8(10):1364–1370.

Xue, M., Atallah, B. V., and Scanziani, M. (2014). Equalizing excitation-inhibition ratios across visual cortical neurons. *Nature*, 511(7511):596–600.

Yang, G. R., Murray, J. D., and Wang, X.-J. (2016). A dendritic disinhibitory circuit mechanism for pathway-specific gating. *Nature Communications*, 7:12815.

Yiu, A. P., Mercaldo, V., Yan, C., Richards, B., Rashid, A. J., Hsiang, H.-L. L., Pressey, J., Mahadevan, V., Tran, M. M., Kushner, S. A., Woodin, M. A., Frankland, P. W., and Josselyn, S. A. (2014). Neurons are recruited to a memory trace based on relative neuronal excitability immediately before training. *Neuron*, 83(3):722–735.

Yizhar, O., Fenno, L. E., Prigge, M., Schneider, F., Davidson, T. J., O'Shea, D. J., Sohal, V. S., Goshen, I., Finkelstein, J., Paz, J. T., Stehfest, K., Fudim, R., Ramakrishnan, C., Huguenard, J. R., Hegemann, P., and Deisseroth, K. (2011). Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature*, 477(7363):171–178.

Zenke, F. and Ganguli, S. (2017). SuperSpike: Supervised learning in multi-layer spiking neural networks. *arXiv preprint arXiv:1705.11146*.

Zhang, L. I., Tan, A. Y. Y., Schreiner, C. E., and Merzenich, M. M. (2003). Topography and synaptic shaping of direction selectivity in primary auditory cortex. *Nature*, 424(6945):201–205.

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figures

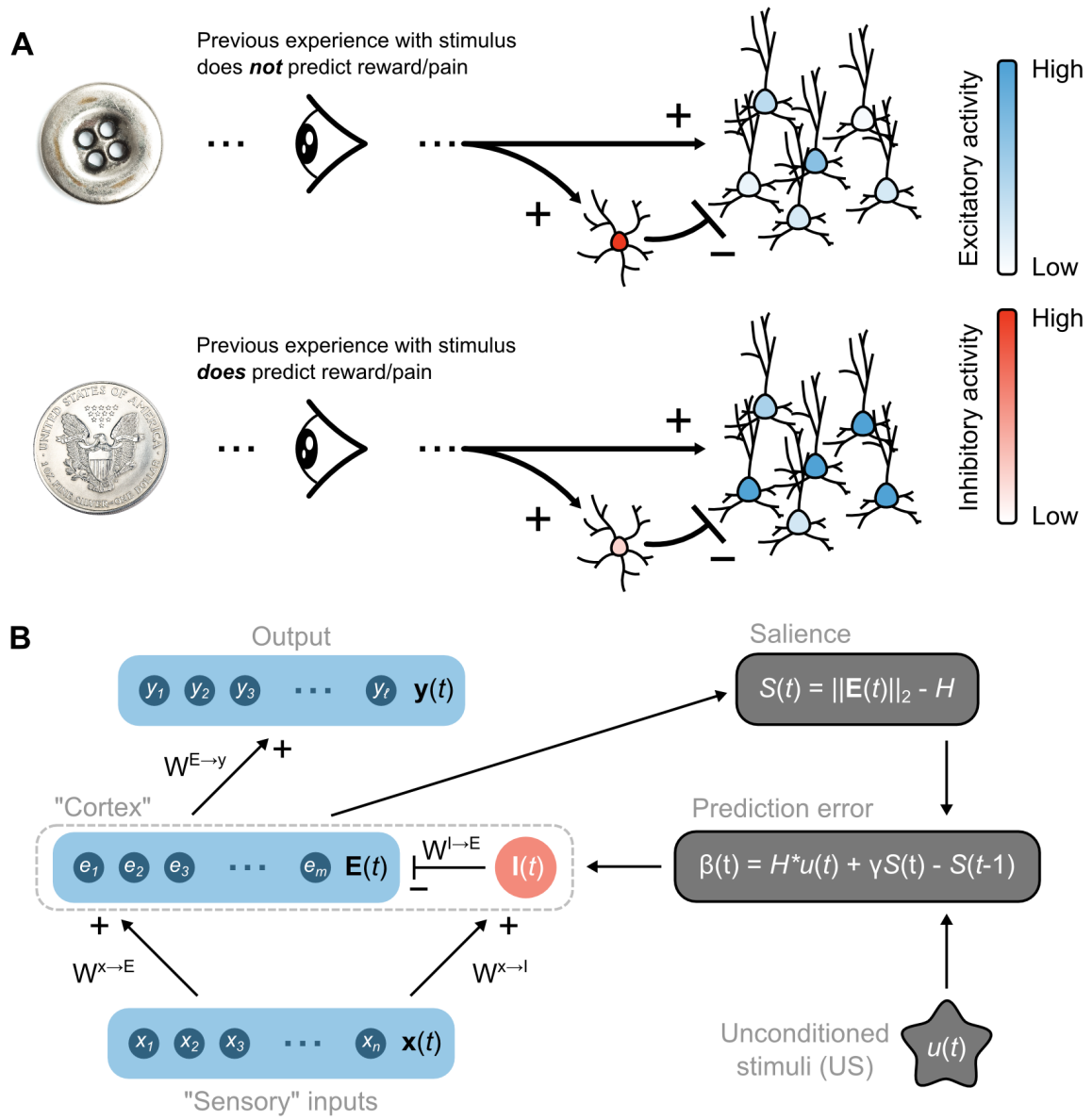

Figure 1. Overview of the proposed relevance code and network model.

(**A**) Schematic illustrating the hypothesis that relevance (prediction of reinforcement) is coded by levels of excitatory neuron output from a network, which is controlled by feed-forward inhibition. (**B**) Basic structure of the network model. Left side shows feed-forward connections from "Sensory" inputs, through inhibitory (I(t)) and excitatory (E(t)) "Cortex" units, with E(t) units feeding onto an output layer. Right-side shows how the salience signal (S(t)), computed from the overall level of excitatory unit activity, is combined with signals about environmental unconditioned stimuli (u(t)) to generate a prediction error that supervises the plasticity of connection weights between Sensory and Cortex layers.

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

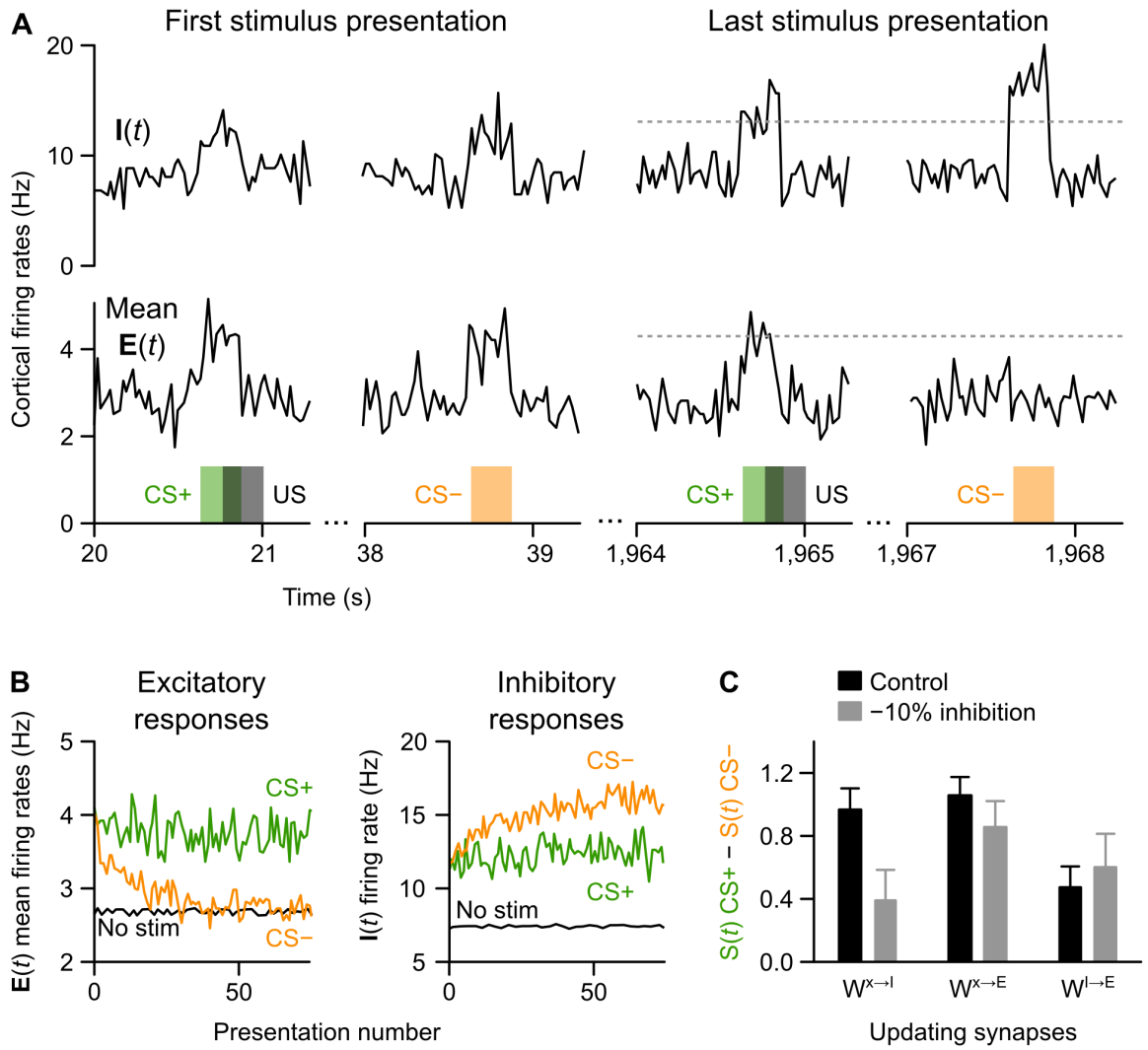

Figure 2. Demonstration of learned irrelevance and its impairment following disruption to inhibition.

(**A**) Average Cortex excitatory unit activity (lower plots) and inhibitory unit activity (upper plots) at simulated, 20 ms time steps in response to unlearned stimuli (left side) compared with the end of a series of repeated presentations (right side). Stimuli are either paired (CS+, red bar) or unpaired (CS-, blue bar) with an unconditioned stimulus (US, grey bar). Excitatory responses were initially high to both stimuli, but after learning they increased only in response to the CS+, demonstrating learned irrelevance for the CS-. (**B**) Averaged excitatory unit (left) and inhibitory unit (right) responses to the CS+ (red) and CS- (blue) across presentations, as compared with non-stimulus periods (grey line). Learned irrelevance took place over the first 20 trials, after which excitatory responses to the CS- plateaued to the same level as non-CS inputs. This was due to increased inhibitory responses to the CS-. (**C**) Excitatory responses to the CS+ relative to the CS- during final presentations are plotted for both control conditions and in simulations of inhibitory dysfunction (means ± STD across 30 model runs). Learned irrelevance was lost with inhibitory neuron disruption only in the inhibitory neuron plasticity model ($W^{\mathbf{x}\to\mathbf{I}}$).

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

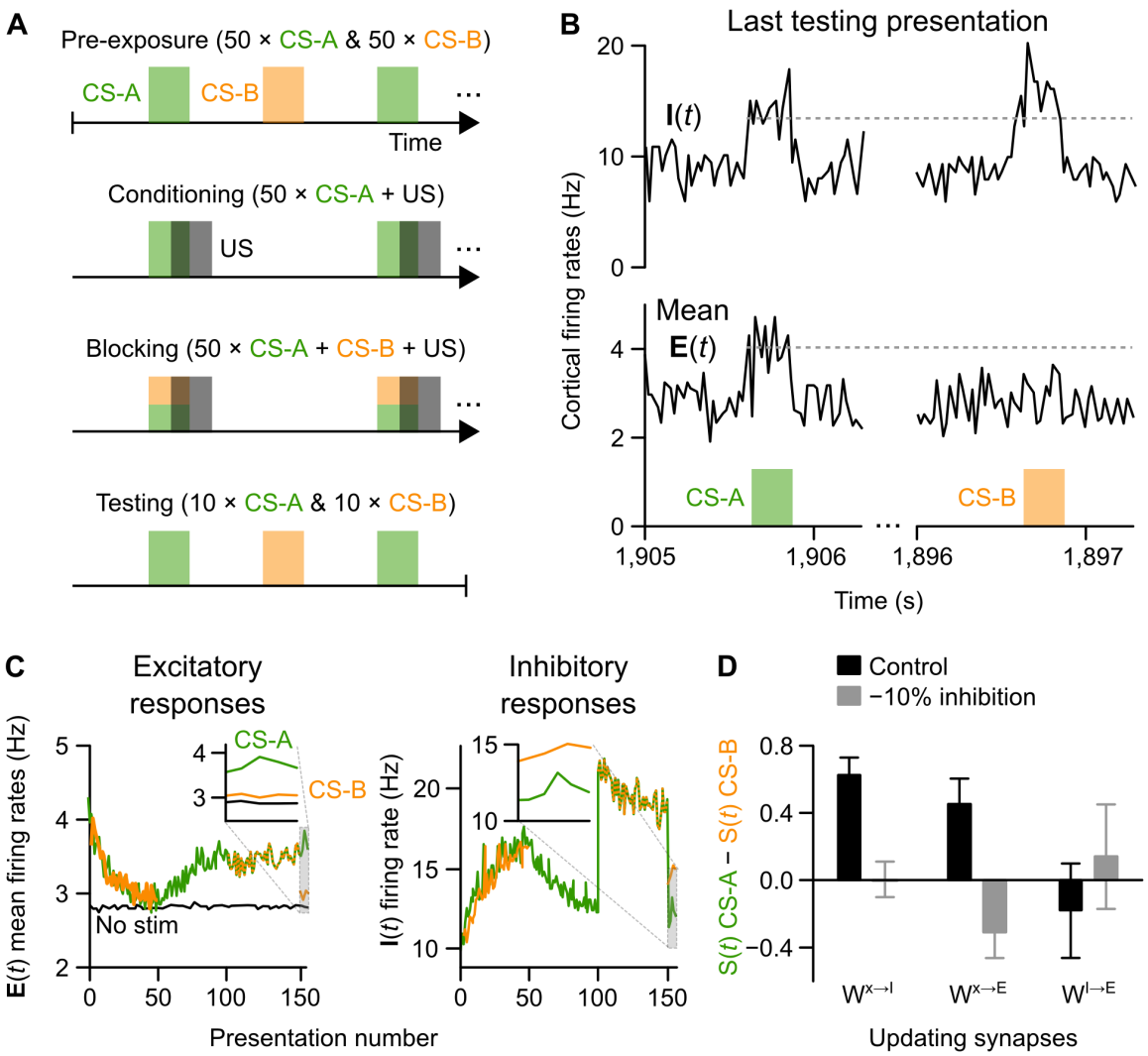

Figure 3. Demonstration of blocking and its impairment following inhibitory disruptions.

(**A**) Illustration of the blocking paradigm: the model was first habituated to two stimuli (CS-A, CS-B; Pre-exposure), the CS-A and a US were then repeatedly presented at partially overlapping times (Conditioning), both CS-A and CS-B were then presented with the US (Blocking), followed by independent presentations of CS-A and CS-B (Testing). (**B**) Excitatory (lower plots) and inhibitory (upper plots) unit activity over 20 ms bins show the networks response to CS-A (left) and CS-B (right) at the end of the blocking paradigm. In spite of CS-B being paired with the US, the "blocked" stimulus did not elicit increased activity among excitatory units. (**C**) Excitatory (left) and inhibitory (right) unit responses to CS-A and CS-B over trials. Test epochs are expanded in insets. (**D**) Excitatory responses to CS-A relative to CS-B at the end of the test epoch are plotted in control simulations and simulations in simulations with dysfunctional inhibition (means ± STD across 30 model runs). The inhibitory neuron plasticity model ($W^{\mathbf{x}\to\mathbf{I}}$) showed a loss of the blocking effect when inhibition was disrupted; unexpectedly, the excitatory neuron plasticity model ($W^{\mathbf{x}\to\mathbf{E}}$) exhibited a reversal of the blocking effect; i.e., CS-B was learned more strongly than CS-A.

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

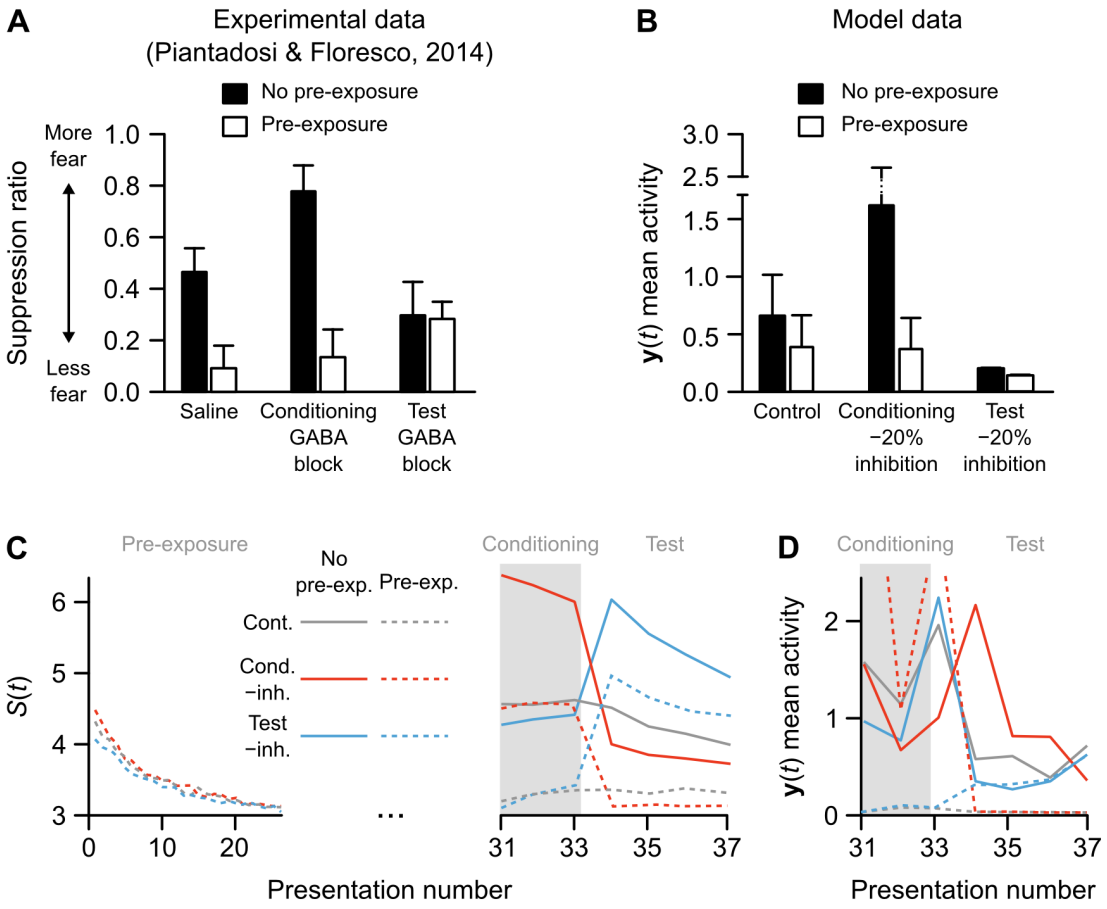

Figure 4. Simulation of experimental data on rodent latent inhibition.

(**A**) Fear expression in rats in a latent inhibition paradigm in which animals were either pre-exposed (black bars) or not pre-exposed (white bars) to the conditioned stimulus, and treated with either saline, a GABA-antagonist during conditioning, or a GABA-antagonist during testing (reproduced from Piantadosi and Floresco (2014)). (**B**) Data from the model simulation of the same latent inhibition paradigm. Bars show median activity across 30 model runs (errorbars are 90% CI generated by bootstrapping 5-sample medians), of the average Amygdala layer activity during the final (test) stimulus presentations. (**C**) Cortex excitatory unit activity in response to stimuli across trials in an example run of the model. The downward curve during the first 30 presentations show learned irrelevance for all simulations in which the model was pre-exposed to the conditioned stimulus. (**D**) Amygdala activity levels in an example run of the model over trials with Conditioning and Testing epochs (as in the right-side panel of part C). Test period activity shows a pre-exposure effect in the control condition (solid versus dashed gray lines). This is amplified when inhibition is disrupted during conditioning (solid versus dashed red lines) but was lost when inhibition was disrupted during test (solid versus dashed blue lines).

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

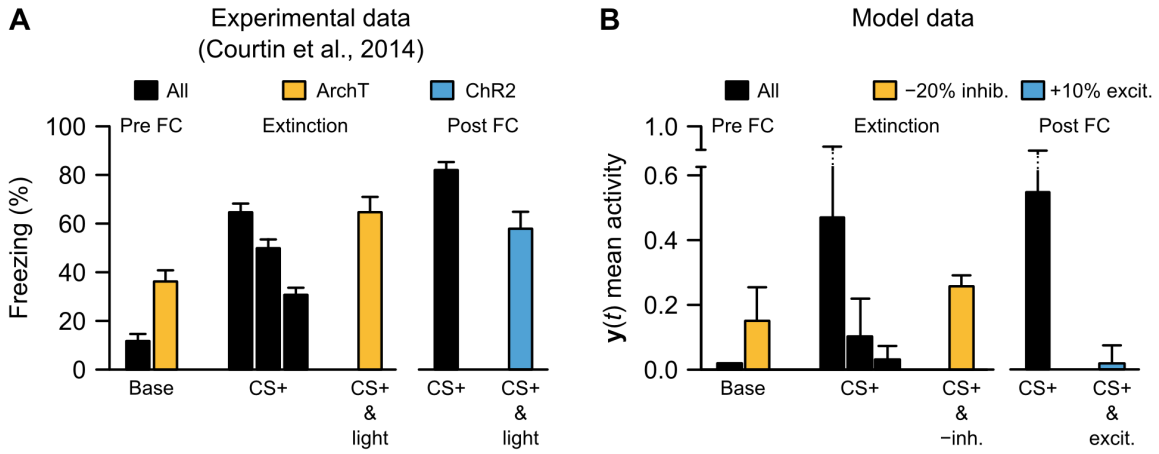

Figure 5. Simulation of experimental data on role of feed-forward inhibition in freezing behavior.

(**A**) Experimental data illustrating the effects of optogenetic inhibition ("ArchT") or excitation ("ChR2") of the medial prefrontal cortex parvalbumin-expressing inhibitory neurons (modified from Courtin et al. (2014)). Inhibitory neuron inhibition was performed both before conditioning ("Base") and following conditioning and extinction (CS+ & light). Inhibitory neuron excitation was performed following conditioning (right side, CS+ & light). (**B**) Replication of general patterns of inhbitory neuron manipulations (part (E)) in the model, substituting -20% inhibition for "ArchT" and +10% excitation (i.e., increased $W^{\mathbf{I}\to\mathbf{E}}$ weights by 10%) for "ChR2".

bioRxiv preprint doi: https://doi.org/10.1101/161331; this version posted July 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

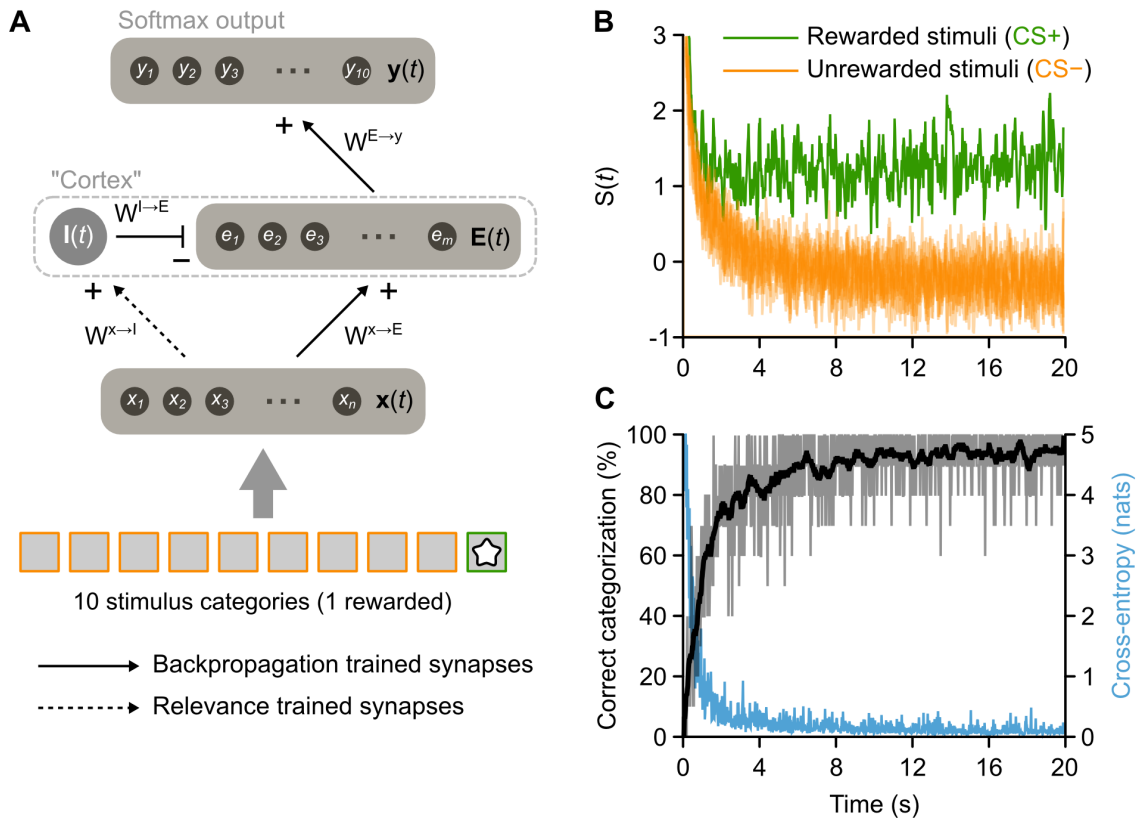

Figure 6. Multiplexed stimulus category and relevance codes via simultaneous excitatory and inhibitory learning.

(**A**) Diagram illustrating modified model that included both the mechanisms described above for relevance learning (on $W^{\mathbf{x}\to\mathbf{I}}$ synapses) in addition to mechanisms learning an output vector that matches categories presented as input (backpropagation algorithm applied to the $W^{\mathbf{E}\to\mathbf{y}}$ and $W^{\mathbf{x}\to\mathbf{E}}$ synapses). As illustrated by bottom boxes, only one of ten stimuli presented to the network was rewarded. (**B**) Average excitatory unit responses to the presented CS+ (red) and the nine CS- stimuli (blue) over trials. As in previous simulations, the network quickly learns respond more strongly to the CS+. (**C**) Performance of the model on input classification. Over the same trials that the network learns to discriminate CS+ and CS- stimuli, it also becomes capable of matching the output vector to the input. Black trace shows percent of presentations that stimuli are correctly classified, which increases quickly before reaching a plateau. Mauve trace shows the cross-entropy, an information measure (in natural units of information) based on the output activity distribution that is inversely related to the strength of input classification.