Inhibitory engrams in perception and memory.

Helen Barron[1,2*], Tim P. Vogels[3*], Timothy Behrens[1,4*,] and Mani Ramaswami[5,6*]

[1]The Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, University of Oxford, Oxford OX3 9DU, UK. [2]Medical Research Council Brain Network Dynamics Unit, Department of Pharmacology, University of Oxford, Oxford OX1 3QT, UK. [3]Centre for Neural Circuits and Behaviour, University of Oxford, Oxford OX1 3SR, UK. [4]The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London WC1N 3BG, UK. [5.]Trinity College Institute of Neuroscience, School of Genetics and Microbiology and School of Natural Sciences/ Trinity College Dublin, Dublin-2 Ireland. [6]National Centre for Biological Sciences, TIFR Centre, Bangalore 560065, India.

Running Title: Excitation Inhibition Balancing in Memory and Cognition

***Correspondence:** helen.barron@merton.ox.ac.uk; behrens@fmrib.ox.ac.uk; tim.vogels@cncb.ox.ac.uk; mani.ramaswami@tcd.ie

Keywords: Neural Representation, Negative Image, Computational, Model, Drosophila, Human, Attention, Neuromodulation, Neural network.

SUMMARY

Nervous systems use excitatory cell assemblies to encode and represent sensory percepts. Similarly, synaptically connected cell assemblies or "engrams" are thought to represent memories of past experience. Multiple lines of recent evidence indicate that brain systems create and use inhibitory replicas of excitatory representations for important cognitive functions. Such matched "inhibitory engrams" can form through homeostatic potentiation of inhibition onto postsynaptic cells that show increased levels of excitation. Inhibitory engrams can reduce behavioral responses to familiar stimuli thereby resulting in behavioral habituation. In addition, by preventing inappropriate activation of excitatory memory engrams, inhibitory engrams can make memories quiescent, stored in a latent form that is available for context-relevant activation. In neural networks with balanced excitatory and inhibitory engrams, the release of innate responses and recall of associative memories can occur through focused disinhibition. Understanding mechanisms that regulate the formation and expression of inhibitory engrams *in vivo* may help not only to explain key features of cognition, but also to provide insight into transdiagnostic traits associated with psychiatric conditions such as autism, schizophrenia and post-traumatic stress disorder (PTSD).

\body

Percepts are thought to be represented in the brain by excitatory activity in groups of neurons, described as cell-assemblies (1). Memories may similarly be represented by excitatory connections across different cell-assemblies, which form when experiences trigger coordinated activity. Several recent studies indicate that the storage and reactivation of these excitatory "engrams" (2) may respectively be accompanied by creation and modulation of matched inhibitory engrams. In this *Perspective*, we propose that inhibitory engrams, also termed "negative images" or "inhibitory representations", explain two fundamental features of animal and human psychology: first, behavioral habituation, which allows organisms to ignore familiar, frequently encountered percepts or stimuli; and second, latent memory, wherein the brain stores tens of thousands of memories in a silent or latent state until recall is required. Such inhibitory engrams can be constructed in neural networks through simple, evolutionarily primitive, synaptic and cellular mechanisms, which are recognized to be involved in the phenomenon of excitatory-inhibitory balance observed across the brain.

Neurons and neural circuits normally operate within a preset range of activity. Outside this range, altered levels of neuronal spiking trigger a variety of homeostatic mechanisms ranging from compensatory ion-channel expression to local synaptic scaling (3, 4). These homeostatic mechanisms ensure that the activity parameters of a neuron operate around a set-point such that the spiking rates remain stable and a balance of depolarizing and hyperpolarizing currents is maintained. This ensures that excitation and inhibition are both locally and globally balanced, despite plastic changes across neurons and synapses.

A potentially important homeostatic mechanism is potentiation of inhibitory synapses, which occurs under specific conditions and acts to prevent excessive levels of postsynaptic activity (5-9). An underappreciated property of this balancing process is that it naturally creates "inhibitory representations" or "negative images" of new, unbalanced excitatory patterns that arise within neural networks in response to experience (10, 11). Here, we discuss how these inhibitory engrams created through compensatory inhibitory potentiation contribute to cognitive function, and ask how their disturbance could contribute to clinical features and transdiagnostic traits observed in neuropsychiatric conditions. In doing so, we integrate insights from diverse neurophysiological, behavioral, computational and brain imaging studies across

multiple species, to extract and highlight fundamental principles of memory consolidation and recall (7, 8, 10-13).

Evidence for restoration of EI balance through inhibitory synapse potentiation.

Excitation and inhibition appear balanced in cortical neurons, both at a global (14-16) and local level (6, 17, 18). At steady state, cortical neurons show closely matched depolarising and hyperpolarising currents (16, 17). During receipt of sensory input the short delay in arrival of matched hyperpolarising currents constrains action potential firing to a brief temporal window (16, 19). However, this balance is disturbed during learning when excitatory plasticity occurs during a transient reduction in inhibition (18) (12, 20, 21). For subsequent stable storage of learned information, the balance between excitation and inhibition must be restored. One means by which this is achieved is via compensatory inhibitory synaptic potentation (6, 22) (Figure 1).

At an electrophysiological level, the process of compensatory inhibitory synaptic potentiation has been best characterized in the mammalian auditory cortex. Principal cells in the auditory cortex normally show balanced depolarising and hyperpolarising currents across a range of tonal frequencies but each cell is optimally tuned to a preferred frequency at which maximal excitatory and inhibitory postsynaptic currents (EPSCs and IPSCs) occur (16). Following a simple form of learning where exposure to a specific tone is paired with direct stimulation of nucleus basalis (NB), a structure that mediates attentional engagement, a shift in the preferred frequency of a principal neuron towards the frequency of the exposed tone can be observed (6) (Figure 1b). In whole cell recordings, this shift is first observed in the excitatory post-synaptic currents (EPSCs), disrupting the balance between excitation and inhibition at the new frequency (6) (Figure 1b). This "representational plasticity" appears to depend upon NB stimulation due to the reduction in feed-forward inhibition following acetlycholine release (21, 23). Strikingly, if tone exposure is continued for 120-180 minutes without associated NB-stimulation, then rebalancing occurs through potentiation of the IPSC, which shifts to match the EPSC with a peak at the new frequency (Figure 1c). The mechanism responsible for this EI-balance restoration has also been investigated using computational approaches.

Computational simulations indicate that simple synaptic plasticity rules are sufficient to account for rebalancing. In a model network of postsynaptic, integrate and fire neurons, Vogels et al. applied a simple spike-time dependent plasticity rule that acts on inhibitory-to-excitatory connections (7). The plasticity rule strengthens inhibitory synapses if inhibitory neurons fire within 80 msec of a postsynaptic spike. When this rule is implemented within an unbalanced system where excitatory synapses are stronger than their inhibitory counterparts, simulations show that inhibitory synaptic potentiation occurs until EPSCs and IPSCs are precisely matched. With some tuning of the target spiking rate for the postsynaptic neuron, the learning rate and the spiking frequency of inhibitory neurons, the experimentally observed phenomenon of EI rebalancing can be accurately reproduced (7) (Figure 1c). Experimental observations and theoretical models therefore agree that inhibitory potentiation plays a critical role in rebalancing cortical networks following excitatory plasticity.

While experimental observations and theoretical models point at a critical role for inhibitory plasticty in rebalancing cortical networks, its biophysical implementation is still poorly understood and underlying mechanisms could vary across brain regions and inhibitory cell types. Moreover, alternative strategies, such as changes in excitatory drive (24), may also contribute to cortical rebalancing. Similarly, at a theoretical level, the precise learning rule responsible for inhibitory potentiation remains open to debate (25) as the implementation of alternative learning rules can also successfully account for cortical rebalancing (26). Nevertheless, additional empirical investigations in rodent auditory cortex emphasise the importance of inhibitory potentiation as a means to rebalance neural circuits, and show that near coincident pre- and post-synaptic activity is required (9). Furthermore, inhibitory potentation is reported to be dependent upon NMDA receptors, suggesting that NMDA receptors act as coincidence detectors to coordinate plasticity between co-active excitatory and inhibitory neurons (27-30).

Negative images in adaptive stimulus filtering and behavioral habituation

Following increased activity across an ensemble of excitatory synapses, the form of EI balancing described above is predicted to result in delayed strengthening of matched inhibitory synapse ensembles (10, 11) (Figure 2A-C). This creates inhibitory engrams or representations, which counterbalance excitatory representations and ensure that EI balance is maintained in face of increased excitation. In large-scale neural networks,

such inhibitory engrams may underlie multiple fundamental cognitive processes. We first consider their role in habituation, a form of non-associative, implicit memory that reduces innate responses to irrelevant stimuli (10, 31, 32).

To explain olfactory habituation in *Drosophila*, behavioral genetic analyses have inferred and invoked an inhibitory learning rule similar to that involved in restoring EI balance in mammalian auditory cortex, (8, 10, 33). In insects, odorants are encoded by assemblies of projection neurons in the antennal lobe, a structure homologous to the mammalian olfactory bulb. Many lines of evidence argue that in a neutral environment, prolonged activation of an odorant-specific excitatory assembly results in the selective strengthening of inhibitory synapses onto neurons activated by the odorant (8, 34). This results in the formation of an inhibitory replica of the specific pattern of odor-induced excitation. The newly created inhibitory representation of odorant-induced excitation acts as a filter to specifically attenuate physiological and behavioural responses to the familar and inconsequential odorant (10). Significantly, as observed for the EI balancing process associated with re-tuning cells in mammalian auditory cortex, insect olfactory habituation also requires postsynaptic NMDA receptors suggesting that a common homeostatic mechanism is at play (8, 9). Together, these observations suggest that olfactory habituation may be usefully considered to arise through a form of EI balancing in which inhibitory potentiation serves to restrain the spiking of a subset of odorant-activated projection neurons to within a behaviorally appropriate range.

Broadly similar mechanisms for auditory habituation have recently been reported in the mammalian cortex, albeit without direct evidence for the underlying synaptic mechanism (13) (Fig 2D-L). Here, *in vivo* GCaMP-based imaging shows that tone-specific auditory habituation is associated with reduced calcium-fluxes in layer 2/3 pyramidal cells in the rodent auditory cortex. This reduction in pyramidal cell activity is accompanied by enhanced activity of somatostatin positive (SOM) inhibitory neurons in the same brain region. Habituation can thus be characterised as a 10-fold reduction in the excitation/inhibition ratio of population activity. Interestingly, when attention to the tone becomes important for task performance, SOM inhibition is reduced and pyramidal cell responses to the tone increase even in habituated animals (13).

Taken together, data from insect and mammalian nervous systems suggest that habituation may generally arise through the formation of inhibitory engrams created via inhibitory synaptic potentiation (6, 7, 9, 10, 13). We note that in the mammalian brain, the subtypes of inhibitory interneurons involved and their mode of regulation may depend on the particular neural circuit in which they are embedded. For instance, while SOM-positive interneurons have been identified as playing a critical role in auditory habituation, PV-positive interneurons are noted for their more pervasive role in equalising the ratio between inhibition and excitation (35). Although the precise interaction between these different subtypes is not yet clear, the model in which matched inhibition drives habituation generates important predictions. One notable corollary suggests that innate behavioural responses attenuated by habituation can later be rapidly restored through disinhibition, via inputs that silence the relevant inhibitory neurons (13).

This provides an explanation for the psychological phenomena of dishabituation and attentional override, two defining properties of habituation (10, 31, 32). Inhibitory engrams can therefore attenuate expression of unnecessary behavioural responses, while selective modulation of EI balance may provide a mechanism to flexibly recover that expression.

Regulating the formation of inhibitory engrams

The formation and affect of inhibitory engrams appears to be regulated by context. By definition, behavioural habituation occurs to non-salient and inconsequential stimuli. Indeed, contextual inputs that confer salience or assign emotional significance to a stimulus are known to actively block habituation (10, 31, 32). This is probably best revealed by physiological studies investigating regulation of intrinsic inhibitory synaptic plasticity (metaplasticity) in the circuit that underlies habituation of a siphon withdrawal reflex in Aplysia (36, 37). When multiple mild stimuli are applied to the tail, a reduced tail-touch induced siphon withdrawal is observed as a consequence of potentiated inhibitory feedback onto siphon motorneurons. By contrast, a single electric shock applied to the tail increases tail-touch induced sensitized siphon withdrawal through serotonin-dependent potentiation of excitatory sensorimotor synapses. Remarkably, the tail-shock induced release of serotonin not only potentiates

excitatory connections but also blocks habituation by reducing the ability of repeated, mild tactile tail stimulation to cause inhibitory interneuron-motorneuron synapse facilitation (36, 37) (Figure 3).

Interestingly, as discussed earlier, in the mammalian auditory cortex inhibitory potentiation following EPSC enhancement occurs in response to continued tonal stimulation (Figure 1C), but this has only been observed in the absence of cholinergic NB stimulation (6). As cholinergic NB afferents are known to mediate disinhibition (6), their activation would be predicted to inhibit inhibitory plasticity required for EI rebalancing. More generally, we speculate that when memories are actively encoded during transient periods of high neuromodulator concentration, EI rebalancing mechanisms may be disrupted. Abnormal persistence of such neuromodulation may result in so-called maladaptive memories that persistently reactivate, as observed in post-traumatic stress disorder (38). In conclusion, processes that underlie the restoration of EI balance in both the retuning of cells in the mammalian auditory cortex and in behavioural habituation best described in invertebrates, not only both rely on inhibitory plasticity, but also appear to be regulated by context-dependent neuromodulation.

Inhibitory engrams in associative memory storage and recall

In addition to the proposed role for inhibitory potentiation in cellular re-tuning and habituation, recent experiments in humans suggest that inhibitory potentiation may also play a fundamental role in the storage and recall of associative memories (11) (Figure 4). Using representational fMRI (39), the relative overlap between representations that support two associated stimuli can be used to index expression of associative memories immediately after learning, within the region of cortex that typically represents the stimuli in question. For example, when participants learn to associate rotationally-invariant abstract shapes, an increase in representational overlap is observed in an anterior region of the lateral occipital complex (LOC) (Figure 1d). This observed increase in representational overlap is thought to reflect excitatory synaptic potentiation that occurs between representations for two associated stimuli during learning (40). However, over time, this representational overlap decreases, rendering cortical memory expression invisible to fMRI 24 hours after learning (Figure 1d). Remarkably, when the level of the cortical inhibitory transmitter GABA is reduced using anodal

transcranial direct current stimulation (tDCS), the expression of the associative memory reappears (Figure 1d). Thus, associative memory transitions from an early form which is visible to fMRI, to a later quiescent state that can only be revealed under conditions of reduced inhibition. This result is consistent with the idea that excitatory synaptic potentiation that occurs during learning, is later matched by equivalent inhibitory synaptic potentiation.

Together these observations point to the following model (Figure 4). Excitatory potentiation during learning enhances connectivity between distinct neuronal ensembles that encode associated stimuli. EI balance is then restored by inhibitory synaptic potentiation, which at a network level creates an inhibitory engram of the newly strengthened excitatory connections to match and cancel/ reduce their effect. This ensures that excitatory connections are balanced by equally strong inhibition. Thus, strong associations are stored in a latent state that allows them to be selectively revealed by disinhibition (11).

These findings are consistent with electrophysiological recordings in a songbird while it listens to a tutor's song (Figure 4F-H) (41). When juvenile zebra finches are exposed to a tutor song, excitatory neurons in HVC are active (Figure 4F). However, in adult birds HVC-neuron responses to tutor song are silenced (Figure 4G). Disinhibiting the HVC via application of a GABA agonist, Gabazine, releases excitation to reveal a response not dissimilar from that observed in developing juvenile birds (Figure 4H) suggesting that they are otherwise silenced by inhibitory inputs. Indeed, as juvenile birds learn to immitate a tutor’s song, inhibitory currents in HVC neurons become more precisely locked to the stimulus. This suggests that inhibition plays an important role in protecting stored representations from interference with closely related information.

The temporal period and the precise mechanisms involved in the inferred EI balancing process are not known. But a particularly attractive notion is that this occurs during sleep, potentially during sharp-wave ripples, when previous experiences and memories are replayed offline (42-44). We speculate that this could be one of the ways in which sleep contributes not only to memory consolidation, but also to homeostatic processes that prepare the brain for new learning.

Within this framework, it remains unproven as to how stored information is normally recalled. While mechanisms for memory recall could vary across brain systems and circuits, we suggest neocortical disinhibition as one potential strategy to allow release of strongly connected excitatory ensembles from balanced strong inhibition (13, 39, 41). Thus, in addition to the established function of local disinhibition to enhance excitatory transmission (18) and initial encoding of memory (12, 20, 21), disinhibition may play a significant role in facilitating release and recall of previously learned, but latent cortical associations (11).

At the microcircuit level, experimental evidence for disinhibition in memory recall is evident following fear conditioning, the most simple form of associative memory (12, 45, 46Courtin, 2014). Here, when a conditioned auditory stimulus (CS) is used to trigger fear-memory recall in rats, synchronous spiking activity of principal neurons is accompanied by phasic inhibition of a subset of parvalbumin-positive (PV) GABAergic interneurons in dorso-medial prefrontal cortex (46). This inhibition of inhibitory neurons can be conceived as a disinhibitory mechanism that promotes release of the fear memory engram and thus recall of the learned fear response. Indeed, optogenetic inhibition of these PV GABAergic interneurons is sufficient to disinhibit prefrontal principal neurons and elicit freezing response (46Courtin, 2014). During natural fear memory recall this disinhibition may be mediated by vasointestinal peptide positive (VIP) interneurons (47-49). However, details of the disinhibitory elements and their circuitry can vary between different brain regions and types of associative learning (45, 50).

The neural mechanisms that mediate disinhibition required for memory recall remain poorly understood. At a cellular level disinhibition could be mediated by neuromodulators, acting for instance through muscarinic acetylcholine receptors on cortical interneurons (51). At a higher level, selective disinhibition may be mediated by neural structures critical for memory recall, such as the hippocampus or prefrontal cortex (PFC). The hippocampus, for example, plays a critical role in resolving contextually distinct but otherwise overlapping memories (52) and a reduction in hippocampal BOLD signal predicts the degree to which intrusive or unwanted memories remain silent (53, 54). Given these observations, we suggest that the hippocampus contributes to selective inhibition of neocortical inhibitory engrams,

providing a means to reactivate otherwise silent memories by selectively modulating EI balance in favor of excitation. The PFC on the other hand, is thought critical for directed forgetting (55) and may exert further top-down control over the hippocampus to prevent involuntary memory recall (53, 56).

Selective disinhibition, via the hippocampus or PFC, may be particularly critical when so called 'inhibitory control' is necessary to prevent recall following exposure to a cue that has the potential to trigger an unwanted memory (55). We further speculate that the precise contribution made by these two brain regions may depend on the age of a memory, with disinhibition for recent and remote memories showing greater dependence upon the hippocampus and PFC respectively (57-59).

A framework for normal and variant memory storage and recall

The observations and arguments above lead to a common systems level framework for memory storage and recall that are built on two strikingly simple principles (Figure 5). (1) Following habituation or associative training respectively, innate behaviors and simple associative memories are masked by inhibitory engrams. These compensatory inhibitory representations ensure that EI balance is maintained despite new learning and may be considered a critical component of memory consolidation. During habituation, inhibitory engrams are created when stimuli are experienced without concurrent engagement of emotional or attentional circuits (Figures 1-2, Figure 5). Following associative learning, we speculate that sharp-wave ripple dependent replay during sleep may play an important role in their formation. (2) Innate behaviours and associative memories can be recalled in appropriate contexts through selective and local disinhibition. We speculate that local disinhibition is driven by contextual or attentional inputs that recruit secondary inhibitory neurons to selectively target cell-types encoding inhibitory engrams (12, 21, 60) (Figure 2; Figure 4; Figure 5). In this manner volitional and directed recall of memories may occur via sensory stimulation or attentional mechanisms, which act to disinhibit and activate cortical subdomains that store specific memories. Although we focus here on inhibitory engrams created by local EI balancing mechanisms, the framework we propose also naturally accommodates inhibitory representations constructed through potential alternative mechanisms (10, 61-65).

In summary, inhibitory engrams provide a homeostatic mechanism that facilitates flexible yet stable storage. This greatly increases the storage capacity of neural networks, protecting against representational interference and runaway excitation. Inhibitory engrams also provide a gating mechanism that ensures information is stored silently and only released or expressed at time points that are relevant for behavior. This framework appears to have considerable explanatory and heuristic value. For instance, it conceptually distinguishes between behavioral amnesia arising from defects in storage from those arising from defects in recall. It is therefore relevant to recent work which shows that a memory engram remains preserved in amnesic mice that do not express contextual fear memory (66, 67). Consistent with the framework we present here, in the mouse models of Alzheimer's disease, memory storage largely remains intact while recall is disrupted, potentially by a failure of disinhibitory pathways that mediate recall of stored memories (Figure 5).

Implications for EI disruption in clinical conditions

Many investigators have worked towards establishing a common framework to account for the behavioural phenotypes common to psychiatric conditions, which otherwise have complex and distinct genetic bases (68-73). Our premise that a major function of EI balancing is to mask irrelevant perceptions, memories and behavior, allows mechanistic predictions to be drawn that concern specific cognitive dysfunctions that may arise when the process is disrupted (Figure 5). These predictions are particularly relevant, given several lines of genetic and physiological evidence pointing to an imbalance between excitation and inhibition as a possible substrate for symptoms observed in a range of clinical conditions including autism and schizophrenia (70, 71, 74-81).

One predicted consequence of defects in the formation of inhibitory engrams is weak behavioral habituation that could arise either from increased neuromodulatory activity which blocks the formation of inhibitory representations, or from defects in inhibitory function and plasticity. Remarkably, weak habituation is a particularly common feature of autism spectrum disorders (ASD) that is also observed in schizophrenia (31, 82-85). Weak habituation could also result in several downstream cognitive changes observed in autism and in schizophrenia, including altered sensory gating, stimulus hypersensitivity and reduced ability to cope in complex environments, where multiple

signals may appear salient and compete for attention (10, 31, 82, 85, 86). It is also conceivable that additional features of autism such as sticky attention, wherein familar stimuli remain engaging for unusually long periods of time while novel stimuli seem challenging, could arise from weak habituation and hypersensitivity to novel stimuli (85-87). An additional predicted consequence of defects in inhibitory rebalancing that we postulate to be necessary for masking associative memories, is strong and contextually inappropriate recall of associative memories. Strikingly, unusually strong and vivid associative memories are often associated with ASD (88). It is conceivable that vivid recall of unconnected memory traces could lead to inappropriate associations and thought disturbances and delusions observed in schizophrenia (Figure 5). This is consistent with theoretical models in which EI imbalance in neural networks has been used to account for hallucinations and delusional experiences that may be observed both in autism and in schizophrenia (86, 89, 90).

This mechanistic proposal that specific cognitive features and transdiagnostic traits of autism and schizophrenia arise in part due to defects in the creation of inhibitory engrams is now experimentally testable, taking advantage of MRI paradigms that provide an index for EI balance across associative memories (11, 39). One simple testable prediction is that EI imbalance across cortical associations will persist for unusally long time periods in affected individuals. While the validity of this proposal will require additional investigation, we suggest that a model that connects synaptic mechanisms of EI balancing in neural circuits with fundamental cognitive processes altered in psychiatric conditions may provide a unifying principle to explain common disturbances that arise from diverse molecular and genetic perturbations.

AKNOWLEDGEMENTS

We thank Jens Hillebrand for help with the figures, Isabell Twick for comments on the manuscript, Kevin Mitchell, Tomas Ryan, Srikanth Ramaswamy and Mriganka Sur for useful discussions. H.C.B. is supported by a Junior Research Fellowship from Merton College, University of Oxford, and the John Fell Oxford University Press (OUP) Research Fund (153/046). T.P.V. is funded by a Sir Henry Dale Fellowship from the Wellcome Trust (WT100000). T.E.J.B. is supported by a Wellcome Trust Senior Research Fellowship (WT104765MA) together with a James S. McDonnell Foundation

Award (JSMF220020372). M.R. acknowledges grants from Science Foundation Ireland and the Simons Foundation Autism Research Initiative as well as core support from the NCBS, Tata Institute of Fundamental Research, Bangalore for collaborative work with K. VijayRaghavan.

REFERENCES

1. Hebb DO (1949) *The Organization of Behavior: A Neuropsychological Theory* (Wiley and Sons., New York).
2. Semon R (1921) Chapter II. Engraphic Action of Stimuli on the Individual. . *The Mneme,* ed Simon TbL (George Allen and Unwin, London).
3. Nelson SB & Turrigiano GG (2008) Strength through diversity. *Neuron* 60(3):477-482.
4. Turrigiano G (2012) Homeostatic synaptic plasticity: local and global mechanisms for stabilizing neuronal function. *Cold Spring Harbor perspectives in biology* 4(1):a005736.
5. Fischer TM & Carew TJ (1993) Activity-dependent potentiation of recurrent inhibition: a mechanism for dynamic gain control in the siphon withdrawal reflex of Aplysia. *J Neurosci* 13(3):1302-1314.
6. Froemke RC, Merzenich MM, & Schreiner CE (2007) A synaptic memory trace for cortical receptive field plasticity. *Nature* 450(7168):425-429.
7. Vogels TP, Sprekeler H, Zenke F, Clopath C, & Gerstner W (2011) Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* 334(6062):1569-1573.
8. Das S, *et al.* (2011) Plasticity of local GABAergic interneurons drives olfactory habituation. *Proc Natl Acad Sci U S A* 108(36):E646-654.
9. D'Amour J A & Froemke RC (2015) Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. *Neuron* 86(2):514-528.
10. Ramaswami M (2014) Network plasticity in adaptive filtering and behavioral habituation. *Neuron* 82(6):1216-1229.
11. Barron HC, *et al.* (2016a) Unmasking Latent Inhibitory Connections in Human Cortex to Reveal Dormant Cortical Memories. *Neuron* 90(1):191-203.
12. Letzkus JJ, Wolff SB, & Luthi A (2015) Disinhibition, a Circuit Mechanism for Associative Learning and Memory. *Neuron* 88(2):264-276.
13. Kato HK, Gillet SN, & Isaacson JS (2015) Flexible Sensory Representations in Auditory Cortex Driven by Behavioral Relevance. *Neuron* 88(5):1027-1039.
14. Moore CI & Nelson SB (1998) Spatio-temporal subthreshold receptive fields in the vibrissa representation of rat primary somatosensory cortex. *J Neurophysiol* 80(6):2882-2892.

15. van Vreeswijk C & Sompolinsky H (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* 274(5293):1724-1726.

16. Wehr M & Zador AM (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* 426(6965):442-446.

17. Okun M & Lampl I (2008) Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature neuroscience* 11(5):535-537.

18. Vogels TP & Abbott LF (2009) Gating multiple signals through detailed balance of excitation and inhibition in spiking networks. *Nature neuroscience* 12(4):483-491.

19. Pouille F & Scanziani M (2001) Enforcement of temporal fidelity in pyramidal cells by somatic feed-forward inhibition. *Science* 293(5532):1159-1163.

20. Kruglikov I & Rudy B (2008) Perisomatic GABA release and thalamocortical integration onto neocortical excitatory cells are regulated by neuromodulators. *Neuron* 58(6):911-924.

21. Chen N, Sugihara H, & Sur M (2015) An acetylcholine-activated microcircuit drives temporal dynamics of cortical activity. *Nature neuroscience* 18(6):892-902.

22. Kuchibhotla KV, *et al.* (2016) Parallel processing by cortical inhibition enables context-dependent behavior. *Nature neuroscience*.

23. Sur M, Nagakura I, Chen N, & Sugihara H (2013) Mechanisms of plasticity in the developing and adult visual cortex. *Prog Brain Res* 207:243-254.

24. Banerjee A, *et al.* (2016) Jointly reduced inhibition and excitation underlies circuit-wide changes in cortical processing in Rett syndrome. *Proc Natl Acad Sci U S A* 113(46):E7287-E7296.

25. Vogels TP, *et al.* (2013) Inhibitory synaptic plasticity: spike timing-dependence and putative network function. *Front Neural Circuits* 7:119.

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

27. Nugent FS & Kauer JA (2008) LTP of GABAergic synapses in the ventral tegmental area and beyond. *The Journal of physiology* 586(6):1487-1493.

28. Nugent FS, Penick EC, & Kauer JA (2007) Opioids block long-term potentiation of inhibitory synapses. *Nature* 446(7139):1086-1090.

29. Sudhakaran IP, *et al.* (2012) Plasticity of recurrent inhibition in the Drosophila antennal lobe. *J Neurosci* 32(21):7225-7231.

30. Woodin MA, Ganguly K, & Poo MM (2003) Coincident pre- and postsynaptic activity modifies GABAergic synapses by postsynaptic changes in Cl- transporter activity. *Neuron* 39(5):807-820.

31. Rankin CH, *et al.* (2009) Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiology of learning and memory* 92(2):135-138.

32. Thompson RF & Spencer WA (1966) Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychological review* 73(1):16-43.

33. Larkin A, *et al.* (2010) Central synaptic mechanisms underlie short-term olfactory habituation in Drosophila larvae. *Learn Mem* 17(12):645-653.

34. Glanzman DL (2011) Olfactory habituation: fresh insights from flies. *Proc Natl Acad Sci U S A* 108(36):14711-14712.

35. Xue M, Atallah BV, & Scanziani M (2014) Equalizing excitation-inhibition ratios across visual cortical neurons. *Nature* 511(7511):596-600.

36. Fischer TM, Blazis DE, Priver NA, & Carew TJ (1997) Metaplasticity at identified inhibitory synapses in Aplysia. *Nature* 389(6653):860-865.

37. Bristol AS & Carew TJ (2005) Differential role of inhibition in habituation of two independent afferent pathways to a common motor output. *Learn Mem* 12(1):52-60.

38. Ehlers A & Clark DM (2000) A cognitive model of posttraumatic stress disorder. *Behaviour research and therapy* 38(4):319-345.

39. Barron HC, Garvert MM, & Behrens TE (2016b) Repetition suppression: a means to index neural representations using BOLD? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 371(1705).

40. Nabavi S, *et al.* (2014) Engineering a memory with LTD and LTP. *Nature* 511(7509):348-352.

41. Vallentin D, Kosche G, Lipkind D, & Long MA (2016) Neural circuits. Inhibition protects acquired song segments during vocal learning in zebra finches. *Science* 351(6270):267-271.

42. Ji D & Wilson MA (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature neuroscience* 10(1):100-107.

43. Eschenko O, Ramadan W, Molle M, Born J, & Sara SJ (2008) Sustained increase in hippocampal sharp-wave ripple activity during slow-wave sleep after learning. *Learn Mem* 15(4):222-228.

44. Ramadan W, Eschenko O, & Sara SJ (2009) Hippocampal sharp wave/ripples during sleep for consolidation of associative memory. *PloS one* 4(8):e6697.

45. Wolff SB, *et al.* (2014) Amygdala interneuron subtypes control fear learning through disinhibition. *Nature* 509(7501):453-458.

46. Courtin J, Karalis N, Gonzalez-Campo C, Wurtz H, & Herry C (2014) Persistence of amygdala gamma oscillations during extinction learning predicts spontaneous fear recovery. *Neurobiology of learning and memory* 113:82-89.

47. David C, Schleicher A, Zuschratter W, & Staiger JF (2007) The innervation of parvalbumin-containing interneurons by VIP-immunopositive interneurons in the primary somatosensory cortex of the adult rat. *Eur J Neurosci* 25(8):2329-2340.

48. Pfeffer CK, Xue M, He M, Huang ZJ, & Scanziani M (2013) Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature neuroscience* 16(8):1068-1076.

49. Pi HJ, *et al.* (2013) Cortical interneurons that specialize in disinhibitory control. *Nature* 503(7477):521-524.

50. Fu Y, *et al.* (2014) A cortical circuit for gain control by behavioral state. *Cell* 156(6):1139-1152.

51. Kim JH, *et al.* (2016) Selectivity of Neuromodulatory Projections from the Basal Forebrain and Locus Ceruleus to Primary Sensory Cortices. *J Neurosci* 36(19):5314-5327.

52. Bulkin DA, Law LM, & Smith DM (2016) Placing memories in context: Hippocampal representations promote retrieval of appropriate memories. *Hippocampus* 26(7):958-971.

53. Benoit RG & Anderson MC (2012) Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron* 76(2):450-460.

54. Depue BE, Curran T, & Banich MT (2007) Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *Science* 317(5835):215-219.

55. Anderson MC, Bunce JG, & Barbas H (2016) Prefrontal-hippocampal pathways underlying inhibitory control over memory. *Neurobiology of learning and memory* 134 Pt A:145-161.

56. Benoit RG, Hulbert JC, Huddleston E, & Anderson MC (2015) Adaptive top-down suppression of hippocampal activity and the purging of intrusive memories from consciousness. *J Cogn Neurosci* 27(1):96-111.

57. Tse D, *et al.* (2011) Schema-dependent gene activation and memory encoding in neocortex. *Science* 333(6044):891-895.

58. Lesburgueres E, *et al.* (2011) Early tagging of cortical networks is required for the formation of enduring associative memory. *Science* 331(6019):924-928.

59. Kitamura T, *et al.* (2017) Engrams and circuits crucial for systems consolidation of a memory. *Science* 356(6333):73-78.

60. Perisse E, *et al.* (2016) Aversive Learning and Appetitive Motivation Toggle Feed-Forward Inhibition in the Drosophila Mushroom Body. *Neuron* 90(5):1086-1099.

61. Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci* 36(3):181-204.

62. Rao RP & Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* 2(1):79-87.

63. Wacongne C, Changeux JP, & Dehaene S (2012) A neuronal model of predictive coding accounting for the mismatch negativity. *J Neurosci* 32(11):3665-3678.

64. Cooke SF, Komorowski RW, Kaplan ES, Gavornik JP, & Bear MF (2015) Visual recognition memory, manifested as long-term habituation, requires synaptic plasticity in V1. *Nature neuroscience* 18(2):262-271.

65. Kogo N & Trengove C (2015) Is predictive coding theory articulated enough to be testable? *Front Comput Neurosci* 9:111.

66. Roy DS, *et al.* (2016) Memory retrieval by activating engram cells in mouse models of early Alzheimer's disease. *Nature* 531(7595):508-512.

67. Ryan TJ, Roy DS, Pignatelli M, Arons A, & Tonegawa S (2015) Memory. Engram cells retain memory under retrograde amnesia. *Science* 348(6238):1007-1013.

68. Sahin M & Sur M (2015) Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders. *Science* 350(6263).

69. Walsh T, *et al.* (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320(5875):539-543.

70. Yizhar O, *et al.* (2011) Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* 477(7363):171-178.

71. Lisman JE, *et al.* (2008) Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. *Trends in neurosciences* 31(5):234-242.

72. Markram H, Rinaldi T, & Markram K (2007) The intense world syndrome--an alternative hypothesis for autism. *Front Neurosci* 1(1):77-96.

73. Oberman LM, *et al.* (2005) EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Brain Res Cogn Brain Res* 24(2):190-198.

74. Rubenstein JL & Merzenich MM (2003) Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes Brain Behav* 2(5):255-267.

75. Gogolla N, *et al.* (2009) Common circuit defect of excitatory-inhibitory balance in mouse models of autism. *Journal of neurodevelopmental disorders* 1(2):172-181.

76. Nelson SB & Valakh V (2015) Excitatory/Inhibitory Balance and Circuit Homeostasis in Autism Spectrum Disorders. *Neuron* 87(4):684-698.

77. Chou YH, *et al.* (2010) Diversity and wiring variability of olfactory local interneurons in the Drosophila antennal lobe. *Nature neuroscience*.

78. Han S, Tai C, Jones CJ, Scheuer T, & Catterall WA (2014) Enhancement of Inhibitory Neurotransmission by GABAA Receptors Having alpha2,3-Subunits Ameliorates Behavioral Deficits in a Mouse Model of Autism. *Neuron* 81(6):1282-1289.

79. Penagarikano O, *et al.* (2011) Absence of CNTNAP2 leads to epilepsy, neuronal migration abnormalities, and core autism-related deficits. *Cell* 147(1):235-246.

80. Tuchman R, Cuccaro M, & Alessandri M (2010) Autism and epilepsy: historical perspective. *Brain Dev* 32(9):709-718.

81. Tyzio R, *et al.* (2014) Oxytocin-mediated GABA inhibition during delivery attenuates autism pathogenesis in rodent offspring. *Science* 343(6171):675-679.

82. Ethridge LE, *et al.* (2016) Reduced habituation of auditory evoked potentials indicate cortical hyper-excitability in Fragile X Syndrome. *Translational psychiatry* 6:e787.

83. Guiraud JA, *et al.* (2011) Differential habituation to repeated sounds in infants at high risk for autism. *Neuroreport* 22(16):845-849.

84. Kleinhans NM, *et al.* (2009) Reduced neural habituation in the amygdala and social impairments in autism spectrum disorders. *Am J Psychiatry* 166(4):467-475.

85. Sinha P, *et al.* (2014) Autism as a disorder of prediction. *Proc Natl Acad Sci U S A* 111(42):15220-15225.

86. Vogels TP & Abbott LF (2007) Gating deficits in model networks: a path to schizophrenia? *Pharmacopsychiatry* 40 Suppl 1:S73-77.

87. Landry R & Bryson SE (2004) Impaired disengagement of attention in young children with autism. *Journal of child psychology and psychiatry, and allied disciplines* 45(6):1115-1122.

88. Zamoscik V, Mier D, Schmidt SN, & Kirsch P (2016) Early Memories of Individuals on the Autism Spectrum Assessed Using Online Self-Reports. *Frontiers in psychiatry* 7:79.

89. Gao R & Penzes P (2015) Common mechanisms of excitatory and inhibitory imbalance in schizophrenia and autism spectrum disorders. *Current molecular medicine* 15(2):146-167.

90. Toal F, *et al.* (2009) Psychosis and autism: magnetic resonance imaging study of brain anatomy. *The British journal of psychiatry : the journal of mental science* 194(5):418-425.

FIGURE LEGENDS.

Figure 1: EI rebalancing observed in rat auditory cortex can be computationally simulated by implementing inhibitory plasticity rules

A. A schematic canonical circuit diagram showing excitatory neurons in red and inhibitory neurons in blue. The postsynaptic excitatory neuron receives balanced excitatory and inhibitory inputs such that the inhibitory and excitatory currents may be considered to be tuned to the same frequency. **B-C**. *Top panel:* schematic circuit diagram showing the effect of plasticity, initially leading to EI imbalance following excitatory plasticity (**B**), and subsequent restoration of balance following inhibitory plasticity (**C**). **B-C.** *Middle panel: In-vivo* whole-cell recording in primary auditory cortex with EPSCs shown in red and IPSCs shown in blue. Concurrent cholinergic stimulation and exposure to a tone (black arrow) that is shifted relative to the original preferred frequency of the neuron (white arrow), modifies the EPSC such that it retunes and peaks at the frequency of the exposed tone (**B**). After repetitive tonal stimulation (in the absence of cholinergic activity) the IPSC eventually shifts from the original to the new frequency to matches and rebalances the modified EPSC (**C**). **B-C.** *Lower panel:* Simulation from a neural network model where the excitatory tuning curve (red circles) is manually changed and a spike-time dependent inhibitory plasticity rule then applied. After 30 minutes of simulation the inhibitory tuning curve (blue squares) still shows the original tuning (**B**), however after 180 minutes it has shifted to match and rebalance the excitatory tuning (**C**). Data panels are adapted from (6, 7).

Figure 2: Inhibitory representations can mediate sensory habituation

A-C. Schematic model for habituation based on (10). **A**. A balanced circuit of inhibitory (blue) neurons and excitatory (red) neurons showing a connected excitatory representation. **B**. Continuous exposure to the relevant percept for the excitatory representation referred to in *A* results in EI imbalance, due to sustained excitatory activity (bright red). **C**. EI rebalancing through inhibitory potentiation creates a matched inhibitory representation (bright blue) to suppress the excitatory representation. **D-I.** Direct evidence for inhibitory potentiation during habituation from mouse auditory cortex (adapted from (13)). **D**. Habituation protocol: mice were passively exposed to a tone for 5 days. Mice experienced 200 trials per day and on each

trial the tone lasted for 5 to 9 seconds. **E-F**. In-vivo two-photon image of tone-evoked GCaMP6s-expressing neurons in layer 2/3 of auditory cortex on day 1 (**E**), and on day 5 (**F**). Excitation was sparser day 5. **G.** Canonical cortical microcircuit showing SOM-positive inhibitory neurons connecting to layer 2/3 (L2/3) excitatory neurons. **H-I.** The daily change index, used to assess the change in activity across days 1-5, showed reduced excitation in layers 2/3 (**H**) which was accompanied by an increase in excitation in SOM-positive inhibitory neurons (**I**). Together these results suggest a reduction in excitation and an increase in inhibition following habituation. **J-L.** In habituated animals, tone-relevant task engagement is accompanied by disinhibition of L2/3 excitatory cells (adapted from (13)). **J**. Tone-relevant task protocol: Habituated animals were trained to lick a food port in response to the habituated tone. The response was then compared to passive exposure to the habituated tone, in the absence of the food port. **K-L.** The change index reflects the increase in the layer 2/3 neuron excitation during the tone-relevant task, relative to the passive exposure. A task associated increase in the excitatory response was observed (**K**), which was accompanied by a decrease in tone-evoked activity in SOM-positive neurons (**L**), consistent with attentional disinhibition.

Figure 3: Neuromodulation gates inhibitory potentiation and EI balancing. (Adapted from (36)).

A. Simplified circuitry that mediates the tail-siphon withdrawal (T-SW) response in *Aplysia*. Sensory neurons (SN) in the tail activate an excitatory interneuron L29 (shown in pink) which excites both the siphon motor neuron (MN) and the inhibitory interneuron L30 (shown in blue). L30 mediates feedback inhibition onto L29. Tail shock results in serotonin release on both L30-L29 and sensorimotor synapses. T-SW habituation is accompanied by enhanced L30-L29 transmission (increased IPSCs recorded in L29), which can also be induced by direct L30 stimulation at frequencies normally induced by tail touch (36, 37). **B-C**. Brief tetanic stimulation was delivered to the inhibitory interneuron L30 (shaded bar indicates stimulation period). The IPSC was recorded from the excitatory interneuron L29. Following stimulation, the IPSC from L29 was enhanced for up to 60 seconds (control condition, open circles). However, this enhancement was not observed if the tetanic stimulation to the inhibitory interneuron

L30 was delivered 90 seconds after a tail-shock (**B**, filled squares, * indicates a significant difference of $p<0.05$ relative to the control condition). Furthermore, this enhancement was also not observed if the tetanic stimulation to the inhibitory interneuron L30 was delivered 90 seconds after application of serotonin (**C**, filled squares, * indicates a significant difference of $p<0.05$ relative to the control condition). Together these results suggest that inhibitory potentiation is similarly blocked by either tail-shock or direct application of serotonin. In other scenarios, neuromodulators could act by prevent inhibitory neuron firing.

Figure 4: Inhibitory rebalancing allows associative information to lie dormant unless EI balance is disturbed

A-C. Schematic representation of a neural network that stores two distinct cell assemblies which are represented by an ensemble of excitatory (either red or orange circles) and inhibitory (either blue or green squares). Initially the two cell assemblies are distinct, with one responding to the red input, and the other responding to the orange input (**A**). When the two cell assemblies become associated via excitatory plasticity between the two excitatory ensembles (black lines), co-activation is observed such that the red cell assembly responds to the orange input and the orange cell assembly responds to the red input (**B**). If inhibitory plasticity then acts to restore balance in the network, the newly strengthened excitatory connections are quenched by inhibition such that co-activation between the two cell assemblies is reduced (**C**). **D-E.** Adapted from (11). **D.** In the human brain, representational fMRI can be used to index co-activation of neural representations for associated abstract shapes in an anterior region of the lateral occipital complex. However, this co-activation is only observed immediately after learning (timepoint 1), and decreases over time (timepoint 2 (~1 hour after learning) and timepoint 3 (~24 hours after learning)). Once the associative memory is quiescent, application of anodal transcranial direct current stimulation leads to a significant reduction in cortical GABA and an increase in the co-activation index for associated stimuli. This suggests that associative memories are stored in cortex in balanced excitatory-inhibitory ensembles, which lie dormant unless the balance between excitation and inhibition is disrupted. **E.** The increase in the co-activation index between associated representations can be predicted by the drop in GABA induced by brain stimulation. **F-H.** Adapted from (41). **F.** In juvenile songbirds,

premotor neurons show an increase in activity during exposure to a tutor song. **G.** In adult songbirds, premotor neurons are quiescent during exposure to a tutor song, a consequence of inhibition selective to song elements that have already been learned. **H.** Application of a GABA agonist, Gabazine, releases excitation in premotor neurons of the songbird, revealing a response not dissimilar from that observed in developing juvenile birds.

Figure 5: Inhibitory engrams in perception and memory: a model.

A. A hypothetical framework showing how inhibitory engrams could form in neural systems illustrated with reference to their influence in perceptual habituation, memory storage and recall. Excitatory inputs onto an array of postsynaptic neurons are shown as positive red peaks and inhibitory inputs as negative blue peaks. In the illustration above, these constitute excitatory and inhibitory engrams respectively.

Before habituation, sensory stimuli activate excitatory perceptual engrams but trigger relatively weak or imprecise inhibition. Repeated excitatory engram stimulation, with minimal attentional or emotional engagement, results in formation of a matched inhibitory engram that reduces the stimulus response and causes behavioral habituation (10). Attentional inputs or dishabituating stimuli can restore the initial stimulus response by promoting disinhibitory inputs that suppress inhibitory engrams (10, 13).

Similarly, memory is first encoded as excitatory engrams. Over time, matched inhibitory engrams form to rebalance the neural network (11, 41). The formation of these inhibitory memory engrams may occur via homeostatic mechanisms similar to those involved in habituation, potentially during replay of excitatory memory engrams. Once formed, inhibitory engrams allow memories to be stored in a dormant form for context-appropriate recall, which we hypothesise to occur through focussed disinhibition.

(Below in green) Defects in inhibitory engram formation or regulation are predicted to cause perceptual and cognitive abnormalities including transdiagnostic traits observed in autism, schizophrenia and post-traumatic stress disorders.