nature neuroscience

Balanced cortical microcircuitry for maintaining information in working memory

Sukbin Lim[1,4] & Mark S Goldman[1–3]

Persistent neural activity in the absence of a stimulus has been identified as a neural correlate of working memory, but how such activity is maintained by neocortical circuits remains unknown. We used a computational approach to show that the inhibitory and excitatory microcircuitry of neocortical memory-storing regions is sufficient to implement a corrective feedback mechanism that enables persistent activity to be maintained stably for prolonged durations. When recurrent excitatory and inhibitory inputs to memory neurons were balanced in strength and offset in time, drifts in activity triggered a corrective signal that counteracted memory decay. Circuits containing this mechanism temporally integrated their inputs, generated the irregular neural firing observed during persistent activity and were robust against common perturbations that severely disrupted previous models of short-term memory storage. These results reveal a mechanism for the accumulation and storage of memories in neocortical circuits based on principles of corrective negative feedback that are widely used in engineering applications.

Working memory on a timescale of seconds is used to hold information in mind during cognitive tasks such as reasoning, learning and comprehension[1]. Over 40 years ago[2], a neural correlate of working memory was identified when the sustained activity of cells of the prefrontal cortex was shown to encode the identity of a remembered stimulus during a memory period. Since this time, such persistent activity has been observed in a wide range of contexts and brain regions[3]. However, the mechanisms by which it is maintained remain poorly understood.

Biophysically, neurons are inherently 'forgetful' as a result of the rapid leakage of currents out of their membranes. Previous theoretical work[3–7] has suggested that this leakage of currents can be offset if memory cells lie within circuits containing positive-feedback loops that precisely replace leaked currents as they are lost (**Fig. 1a**). Models based on this principle can maintain arbitrarily finely graded levels of persistent activity that, in theory, can last indefinitely. However, if the strengths of the positive-feedback loops are slightly too strong or too weak, activity quickly spirals upward or downward until it either saturates or comes to rest at a baseline level[6,7] (**Fig. 1a**). As a result, positive-feedback models of graded persistent activity require a fine tuning of the level of feedback and are highly sensitive to common perturbations, such as global changes in neuronal or synaptic excitabilities, that disrupt this tuning.

Anatomically, neocortical circuits exhibit a plethora of both positive- and negative-feedback pathways. Although positive feedback has been studied in detail, negative-feedback pathways have received relatively little attention in models of working memory. Inhibition typically has been arranged either in 'double-negative' loops that mediate a disinhibitory form of positive feedback[8] or has served as a global, normalizing background[9]. Here, we suggest that inhibition is critical for providing corrective negative feedback that stabilizes persistent activity.

Our model depends on two primary observations. First, cortical neurons receive massive amounts of both excitation and inhibition that, in a wide range of conditions and brain areas, are believed to be closely balanced[10]. Second, recent studies of frontal cortical circuits have reported differential kinetics in the excitatory pathways onto excitatory versus inhibitory neurons. Excitatory to excitatory connections, commonly associated with positive feedback, have relatively slow kinetics resulting from an abundance of slow NMDA conductances[11–14]. Excitatory to inhibitory connections, which are necessary to drive negative feedback, are relatively fast. We found that these two observations naturally lead to a corrective, negative-derivative form of feedback that counteracts drift in persistent activity.

We examined the basic mechanism by which negative-derivative feedback can contribute to persistent activity and temporal integration and constructed network models based on this mechanism. The resulting derivative-feedback models are more robust to many commonly studied perturbations than previous models that are based purely on positive feedback and, as a result of their inherent balance of inhibition and excitation, produce the highly irregular firing typical of neocortical neuron responses[15,16]. The experimental predictions resulting from our model differentiate it from common positive feedback models. We also discuss implications of our model for the NMDA hypothesis of working memory generation.

RESULTS

Error correction through negative-derivative feedback

In the following, we show how observed features of frontal cortical circuits[11–14,17,18] lead to a mechanism of memory storage based on basic principles of engineering feedback control. In systems using feedback control, a corrective signal is generated to oppose errors whenever a deviation from desired behavior is sensed. For the maintenance

[1]Center for Neuroscience, University of California, Davis, Davis, California, USA. [2]Department of Neurobiology, Physiology and Behavior, University of California, Davis, Davis, California, USA. [3]Department of Ophthalmology and Visual Science, University of California, Davis, Davis, California, USA. [4]Present address: Department of Neurobiology, University of Chicago, Chicago, Illinois, USA. Correspondence should be addressed to M.S.G. (msgoldman@ucdavis.edu).

Received 27 March; accepted 15 June; published online 18 August 2013; doi:10.1038/nn.3492

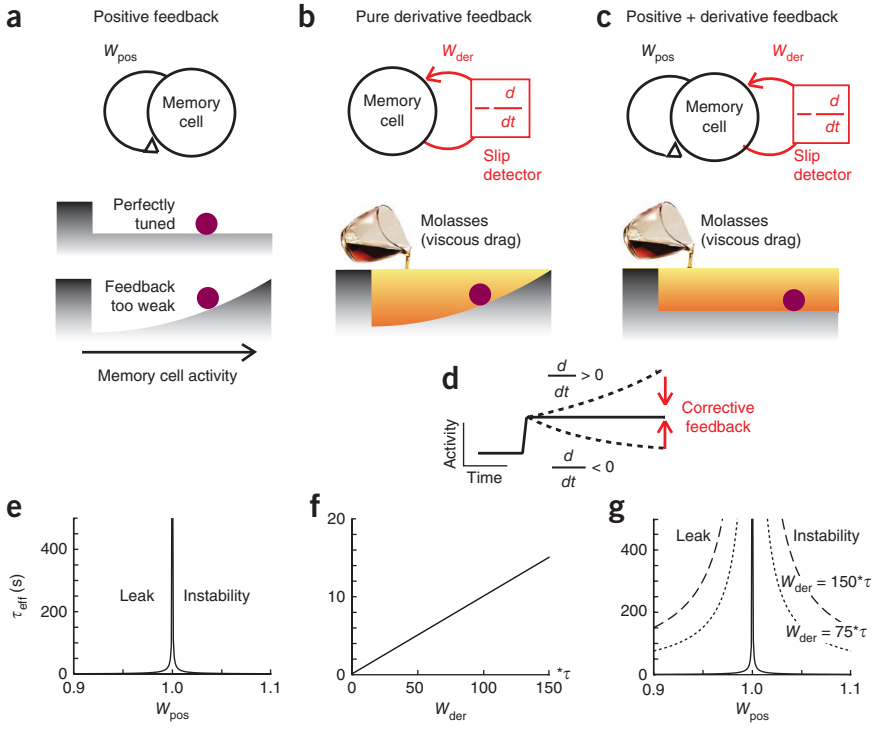

Figure 1 Memory networks with negative-derivative feedback. (**a**–**c**) Simple models of a neural population and their energy surfaces with positive feedback (**a**), derivative feedback (**b**), and hybrid positive and derivative feedback (**c**). Persistent activity can be maintained at different levels (horizontal axis of energy surface) either by a positive-feedback mechanism that effectively flattens the energy surface (**a**,**c**, bottom) or by a negative derivative–feedback mechanism that acts like a viscous drag force opposing changes in memory activity (**b**,**c**, bottom). The wall at the left of the energy surface represents the constraint that activity cannot be negative. (**d**) Illustration of how a negative derivative–feedback mechanism detects and corrects deviations from persistent activity. (**e**–**g**) Effective time constant of activity from equation (2) as a function of the strengths of positive feedback W_{pos} (**e**,**g**) and derivative feedback W_{der} (**f**,**g**). As W_{der} increases, the network time constant τ_{eff} becomes less sensitive to changes in W_{pos} (**g**).

of persistent activity in memory circuits, the deviation to be detected and corrected is a change in time of the memory-storing activity, that is, a temporal derivative (**Fig. 1b**–**d**). If memory activity drifts upward, corresponding to a positive derivative of activity, net inhibition should be provided to reduce the magnitude of this drift. Similarly, if memory activity drifts downwards, net excitation should be increased to offset this drift. In both cases, the required form of corrective feedback is therefore in a direction opposite to the derivative of the neural activity and describes negative-derivative feedback.

To gain a quantitative understanding of how the derivative-feedback mechanism compares to the traditional positive-feedback mechanism, we first considered a simple mathematical model of a memory cell with intrinsic time constant τ that receives a transient input $I(t)$ to be stored in memory (equation (1)). To successfully remember this input after its offset, the memory cell should exhibit only very slow changes dr/dt in its firing rate $r(t)$. This requires that its intrinsic leakage of currents, represented by $-r$, be offset by positive feedback of strength W_{pos} (**Fig. 1a**,**c**) and/or by negative-derivative feedback of strength W_{der} (**Fig. 1b**,**c**)

$$\begin{aligned} \tau\frac{dr}{dt} &= -r + W_{pos}r - W_{der}\frac{dr}{dt} + I(t) \\ \Rightarrow (\tau + W_{der})\frac{dr}{dt} &= -(1 - W_{pos})r + I(t) \end{aligned} \quad (1)$$

Positive-feedback models do not contain the W_{der} term. They maintain persistent firing by providing a feedback current that, when properly tuned by setting $W_{pos} = 1$, offsets the intrinsic tendency of currents to leak out of the membrane. However, if the feedback is too weak ($W_{pos} < 1$), memory activity decays to a baseline level in a manner analogous to an inertia-less particle drifting toward the bottom of a hill (**Fig. 1a**). Similarly, if feedback is too large ($W_{pos} > 1$), activity grows exponentially on a timescale set by the intrinsic time constant τ. Thus, to perform correctly, positive-feedback models require fine tuning of the strength of the positive feedback. Quantitatively, this fine tuning condition is defined by the relation $\tau_{eff} = \tau/(1 - W_{pos})$, where τ_{eff} is the exponential decay time constant of network activity in the presence of positive feedback (equation (1); **Fig. 1e**).

Negative derivative–feedback networks instead slow memory decay by providing a force that opposes the drift of memory activity in a manner mathematically identical to viscous drag forces in fluid mechanics (**Fig. 1b**). This drag force effectively extends the time constant of memory decay in proportion to the strength of the derivative-feedback pathway. For the case in which there is no positive feedback ($W_{pos} = 0$), this leads to an effective network decay time constant $\tau_{eff} = \tau + W_{der}$ (equation (1); **Fig. 1f**).

More generally, negative-derivative feedback can complement positive feedback by opposing drifts that result from imperfect tuning of positive feedback (**Fig. 1c**). In this case, the network time constant (**Fig. 1g**) reflects the effects of both positive- and negative-derivative feedback and, from equation (1), is given quantitatively by

$$\tau_{eff} = (\tau + W_{der})/(1 - W_{pos}) \quad (2)$$

As the negative-derivative feedback gets stronger (contours of increasing W_{der}; **Fig. 1g**), the system becomes increasingly robust to mistuning of the positive feedback W_{pos}. We refer to any network containing a strong negative-derivative feedback component (as in **Fig. 1b**,**c**) as a negative-derivative feedback network. The special subclass of negative-derivative feedback networks with no positive feedback ($W_{pos} = 0$) are denoted as purely negative-derivative feedback networks, whereas those that contain tuned positive feedback ($W_{pos} = 1$) are denoted as hybrid positive and negative derivative–feedback networks.

Negative-derivative feedback in neocortical microcircuitry

How can negative-derivative feedback arise from interactions between excitatory and inhibitory neurons in neocortical circuits? Mathematically, temporal derivatives are created when a signal is subtracted from the same signal offset in time. Similarly, derivative feedback can be created in memory networks by feeding back a memory-storing signal through positive- and negative-feedback pathways that are equal in strength, but have different kinetics. When memory

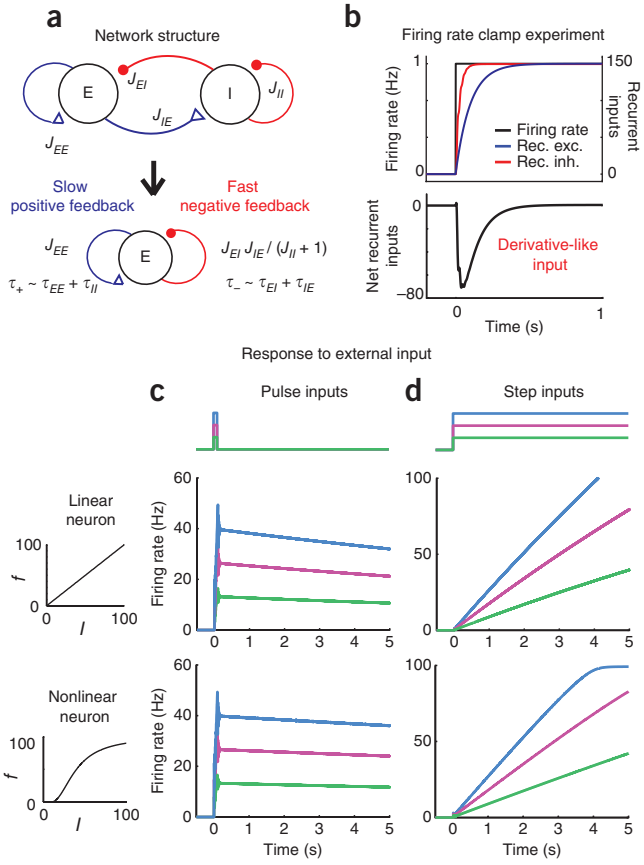

Figure 2 Negative derivative–feedback networks of excitatory and inhibitory populations. (**a**) Derivative-feedback network structure (top) and component feedback pathways onto the excitatory population (bottom). (**b**) In response to external input that steps the excitatory population between two fixed levels, the recurrent feedback pathways mediate a derivative-like signal resulting from recurrent excitation and inhibition that arrive with equal strength, but different timing. (**c**,**d**) Maintenance of graded persistent firing in response to transient inputs (**c**) and integration of step-like inputs into ramping outputs (**d**) with linear (top) and nonlinear (bottom) firing rate (f) versus input current (I) relationships.

activity slips, fast negative feedback mediated by recurrent inhibition rapidly opposes this slip, and slower positive feedback restores the original balance of excitation and inhibition in the circuit. The net effect of this fast inhibition and slow excitation is a feedback signal that opposes changes, that is, generates a negative temporal derivative, of memory cell activity (**Fig. 1b**).

To determine how negative-derivative feedback can arise in a neural network, we constructed a two-population memory circuit model consisting of excitatory (E) and inhibitory (I) populations. The populations were reciprocally connected by synapses of strength J_{ij} and time constant τ_{ij}, where $j = E$ or I denotes the presynaptic population and i denotes the postsynaptic population (**Fig. 2a**). This architecture contains a positive-feedback loop represented by the excitatory-to-excitatory connection of strength J_{EE} and a negative-feedback loop of strength $J_{EI}J_{IE}/(1+J_{II})$ mediated by the excitatory-to-inhibitory-to-excitatory pathway and modulated in strength by the inhibitory-to-inhibitory connection (**Fig. 2a**).

Mathematical analysis of this network to determine the conditions under which persistent activity could be stably maintained revealed two classes of solutions (**Supplementary Modeling**). The first class corresponded to the positive-feedback mechanism ($W_{\text{pos}} = 1$ in equation (1)) and was characterized by having a stronger positive-feedback pathway than negative-feedback pathway so that the net feedback offset the intrinsic leakiness of the neurons. The second class corresponded to negative-derivative feedback, as expressed mathematically by the conditions (see **Supplementary Modeling** for additional inequalities required to maintain network stability)

$$\frac{J_{EE}J_{II}}{J_{EI}J_{IE}} \sim 1 \quad \text{for large } J \text{ values} \tag{3}$$

$$\tau_+ = (\tau_{EE} + \tau_{II}) > (\tau_{EI} + \tau_{IE}) = \tau_- \tag{4}$$

Equation (3) expresses the condition for balancing positive feedback and negative feedback in strength. Equation (4) ensures that the combination τ_+ of synaptic decay time constants associated with positive feedback is slower than the combination τ_- associated with negative feedback; here, τ_{II} acts like a positive feedback contribution because it governs the reduction of negative feedback. Together, equations (3) and (4) define the conditions for negative derivative–like feedback. Strictly speaking, the derivative-like behavior is only at low frequencies, as high frequencies are low-pass filtered by the synapses (**Supplementary Modeling**). This may be advantageous compared with a true derivative, which amplifies high-frequency noise.

To illustrate this derivative-like feedback, we ran a simulation in which the firing rate of the excitatory neuron was clamped by external current injection to go through a perfect step from one steady firing rate to another (**Fig. 2b**). During the periods of steady persistent firing before or long after the step in firing rate, excitation (**Fig. 2b**) and inhibition (red) were balanced, and the net recurrent synaptic input was zero. However, if activity fluctuated, then the different kinetics of the positive- and negative-feedback pathways led to a large, derivative-like recurrent input that opposed the change in network activity (**Fig. 2b**).

Both of the conditions for negative-derivative feedback are present in cortical memory networks. A balance between strong excitatory and inhibitory synaptic inputs has been observed under a wide range of conditions[10], including during sustained activity in prefrontal cortex[17,18]. Slow excitatory-to-excitatory synaptic kinetics have been found that are a result of a prominence of slow NMDA-type receptors[11–14]. When we incorporated these findings into the model, the network maintained long-lasting persistent activity that reflected the level of its transient input (**Fig. 2c** and **Supplementary Fig. 1f**). The network time constant of activity decay, τ_{network}, increased linearly with the J values and with the difference between the time constants τ_+ and τ_-, allowing us to directly connect the network parameters to the strength of derivative feedback in the simpler model of equation (1) through the relation $W_{\text{der}} \sim \tau_{\text{network}} \sim J(\tau_+ - \tau_-)$ (see below and **Supplementary Modeling**). More generally, the network acted as an integrator of its inputs with this same time constant, for example, converting steps of input into linearly ramping activity (**Fig. 2d** and **Supplementary Fig. 1i**).

A potential concern is that the opposition to firing rate changes provided by the negative-derivative feedback mechanism might keep the network from responding to external inputs. However, external inputs comparable to the recurrent inputs in strength, as would be expected if the strengths of both recurrent and external inputs scale with population size, can overcome the derivative feedback and transiently imbalance excitation and inhibition, as observed experimentally during transitions between different levels of sustained activity[17,18] (**Supplementary Modeling**). Furthermore, appropriate arrangement of the external inputs can

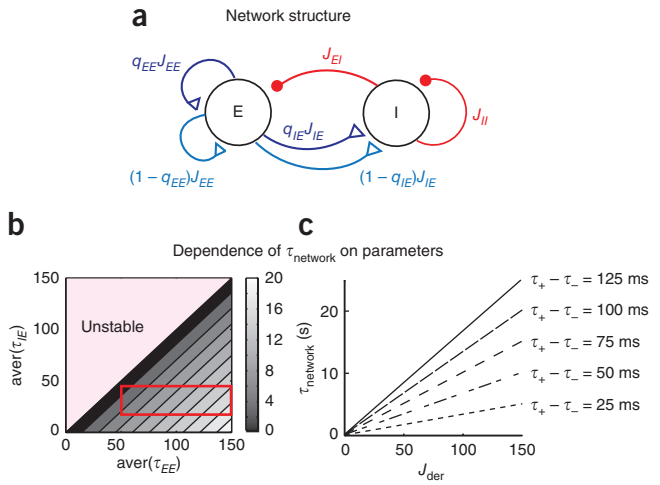

Figure 3 Negative-derivative feedback with mixture of NMDA and AMPA synapses in all excitatory pathways. (**a**) Derivative feedback network structure. Blue, cyan and red curves represent NMDA-mediated, AMPA-mediated and GABA-mediated currents, respectively. q_{EE} and q_{IE} are the fractions of NMDA-mediated synaptic inputs in each excitatory pathway. (**b**) Time constant of decay of network activity, $\tau_{network}$, as a function of the average time constants of excitatory connections, aver(τ_{EE}) and aver(τ_{IE}). Each average time constant is varied either by varying the fractions or by varying the time constants of NMDA-mediated synaptic inputs in each connection. The region in the red rectangle corresponds to a set of possible aver(τ_{EE}) and aver(τ_{IE}) obtained when varying q_{EE} and q_{IE} and holding the synaptic time constants fixed at values matching the experimental observations in ref. 13. (**c**) Time constant of decay of network activity $\tau_{network}$ as a function of the connectivity strengths J_{ij} and the time constants of positive and negative feedback, τ_+ and τ_-. $\tau_{network}$ increases linearly with the balanced amount of positive and negative-derivative feedback $J_{der} \sim J_{EE} \sim J_{IE}J_{EI}/J_{II}$, and with the difference between τ_+ and τ_-, as $W_{der} \sim J_{der}(\tau_+ - \tau_-)$.

reduce the derivative feedback by amplifying this transient imbalance (**Supplementary Modeling**)[19].

Reinterpretation of the NMDA hypothesis for working memory

In traditional positive-feedback models[4,5,20,21], NMDA-mediated synaptic currents computationally serve to provide a nonspecific, slow kinetics process in all feedback pathways. Consistent with this role, NMDA-mediated currents in such models are typically present equally in all neurons, both excitatory and inhibitory. Our model suggests an additional role for NMDA-mediated currents in providing the slow positive-feedback component of a derivative-feedback signal. This requires that the contribution of NMDA-mediated currents be stronger in positive-feedback than in negative-feedback pathways.

To investigate this revised NMDA-hypothesis for memory circuits, we extended our network models to include both NMDA-mediated and non-NMDA (AMPA mediated) currents at all excitatory synapses (**Fig. 3a**). Experimentally, recent measurements of the AMPA- and NMDA-driven components of excitatory transmission have identified two means by which NMDA may contribute more strongly to positive-feedback than to negative-feedback pathways. First, NMDA-mediated currents can be a higher fraction of total excitatory synaptic currents in excitatory-to-excitatory than excitatory-to-inhibitory connections[11,13]. Second, the NMDA-driven component can have slower kinetics[11–14] in excitatory neurons than inhibitory neurons.

We examined quantitatively how this asymmetry in excitatory time constants contributes to negative-derivative feedback. The model with multiple components of excitatory transmission is shown in **Figure 3a**. All excitatory synapses contained both NMDA- and AMPA-type synapses so that both the positive- and negative-feedback loops contained slow and fast synaptic components. Nevertheless, we found that the conditions for derivative feedback–mediated persistent activity followed the same principles identified in the simple network model (**Fig. 2**), that is, a balance between the total positive and negative feedback in strength and slower positive feedback on average. More precisely, the conditions for negative-derivative feedback are still represented by equations of the form of equations (3) and (4). However, J_{EE} and J_{IE} in equation (3) represented the sum of the strengths of NMDA- and AMPA-mediated synaptic currents onto excitatory and inhibitory neurons, respectively, and the time constants τ_+ and τ_- of positive and negative feedback in equation (4) represented the weighted average of the synaptic time constants contributing to positive and negative feedback, respectively (Online Methods and **Supplementary Modeling**).

Thus, even in the presence of slow kinetics in the negative-feedback (excitatory to inhibitory) pathway or fast kinetics in the positive-feedback (excitatory to excitatory) pathway, negative-derivative feedback arises when the positive feedback is slower than the negative feedback on average. As in the simpler networks, the time constant of decay of network activity increased with the difference between the average time constants of positive and negative feedback (**Fig. 3b**,**c**). This slower positive than negative feedback can be achieved either with a higher fraction of NMDA-mediated currents ($q_{EE} > q_{IE}$; **Supplementary Fig. 2a**) or with slower NMDA kinetics ($\tau_{EE}^N > \tau_{IE}^N$; **Supplementary Fig. 2b**) in the excitatory-to-excitatory connection. Thus, this work suggests a revised NMDA hypothesis that highlights the experimentally observed[11–14] asymmetric contribution of NMDA receptors in positive- and negative-feedback pathways as a basis for negative-derivative feedback control.

Robustness of memory performance to common perturbations

A prominent issue in models of neural integration and graded persistent activity is their requirement for tuning of network connection strengths and lack of robustness to perturbations that disrupt this tuning. Several biologically motivated solutions have been proposed to mitigate this problem; for example, a large body of work has shown that the tuning requirements can be greatly reduced if network feedback mechanisms are complemented by cellular[22–24] or synaptic[25–27] persistence mechanisms. However, a largely neglected question in these discussions is whether biological systems are designed to be robust against all types of perturbations and, if not, what types of circuit architectures are robust against the most commonly experienced perturbations.

In traditional positive-feedback models of analog working memory and neural integration, both inhibition (through disinhibitory loops) and excitation mediate positive feedback (see below). As a result, many natural perturbations, including loss of cells, change in cell excitabilities, or changes in the strengths of excitatory or inhibitory synaptic transmission, changed the net level of positive feedback in the network and grossly disrupted persistent firing (**Fig. 4a–f**). In contrast, in models based on derivative feedback (**Fig. 4g–l**), each of these natural perturbations led to offsetting changes. For example, because excitatory cells drive both positive feedback (through excitatory-to-excitatory connections) and negative feedback (through excitatory-to-inhibitory connections), loss of excitatory cells or decrease of excitatory synaptic transmission did not disrupt the balance of positive and negative feedback underlying derivative feedback (**Fig. 4j**). Similarly, changes

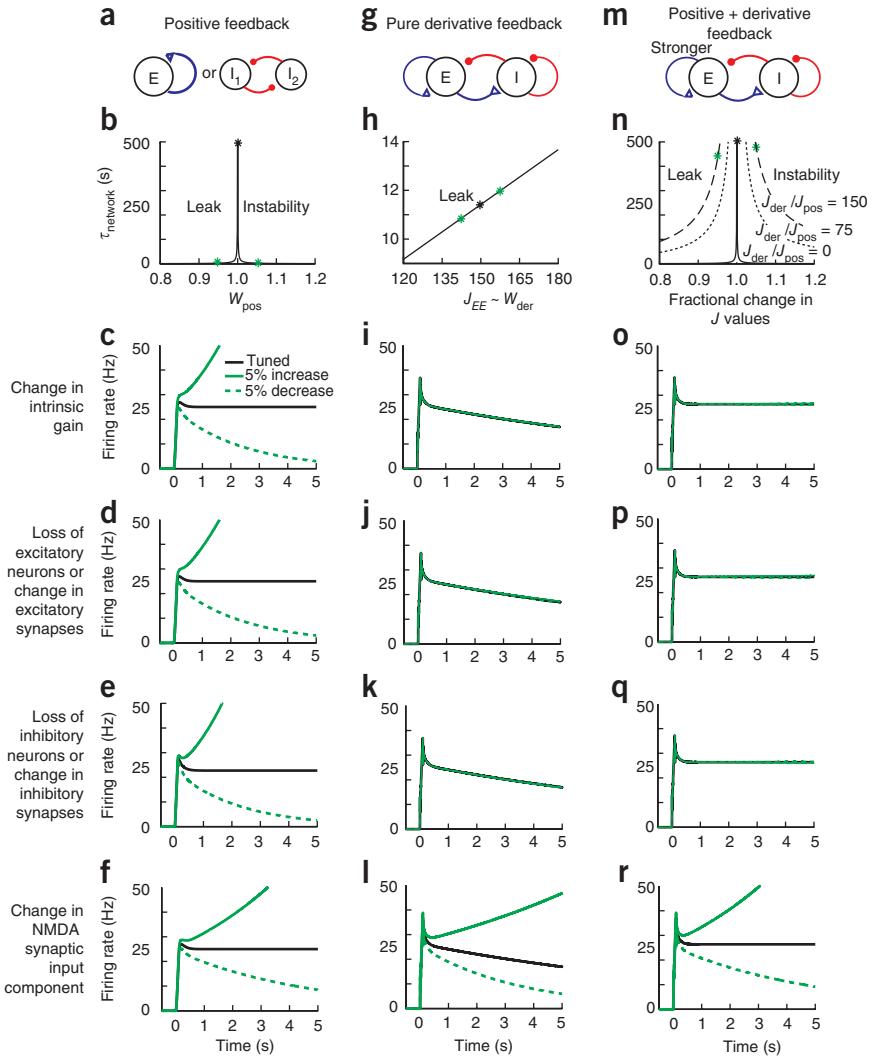

Figure 4 Robustness to common perturbations in memory networks with derivative feedback. (**a**–**f**) Non-robustness of persistent activity in positive-feedback models. (**a**) Positive-feedback models with recurrent excitatory (left) or disinhibitory (right) feedback loops. (**b**) Effective time constant of network activity, $\tau_{network}$, as a function of connectivity strength. Green asterisks correspond to 5% deviations from perfect tuning. (**c**–**f**) Time course of activity in perfectly tuned networks (black) and following small perturbations of intrinsic neuronal gains (**c**) or synaptic connection strengths (**d**–**f**). (**g**–**k**) Robust persistent firing in derivative-feedback models. To clearly distinguish the hybrid models with derivative and positive feedback, purely negative derivative–feedback models with no positive feedback are shown. All excitatory synapses are mediated by both NMDA and AMPA receptors as in **Figure 3**, with parameters chosen to coincide with experimental observations[13]. (**h**) $\tau_{network}$ increases linearly with the strength of recurrent feedback J. (**i**–**k**) Robustness to 5% changes (green asterisks in **h**) in neuronal gains or synaptic connection strengths. (**l**) Disruption of persistent activity in derivative-feedback models following perturbations of NMDA-mediated synaptic currents. (**m**) Hybrid model with positive and derivative feedback. (**n**–**q**) As the strength of negative-derivative feedback is increased, $\tau_{network}$ decreases less rapidly with mistuning than in purely positive-feedback models (**n**) and the network becomes robust against perturbations (**o**–**q**, shown for $J_{der}/J_{pos} = 150$). (**r**) Disruption of persistent activity in the hybrid model following perturbations of NMDA-mediated currents.

in intrinsic neuronal gains did not imbalance the positive and negative feedback received by cells (**Fig. 4i** and **Supplementary Fig. 1**) and changes in inhibitory synapses or loss of inhibitory neurons produced offsetting changes in positive- (inhibitory to inhibitory) and negative-feedback (inhibitory to excitatory) pathways (**Fig. 4k**). Mathematically, the origin of this robustness is that the tuning condition for the derivative-feedback networks (equation (3)) is ratiometric, with the excitation and inhibition received by and projected by a cell population appearing in both the numerator (positive-feedback contributions) and denominator (negative-feedback contributions).

The negative derivative–feedback models are not robust against perturbations that break the balance of inhibition and excitation. For instance, perturbations that differentially affect excitatory-to-excitatory versus excitatory-to-inhibitory synaptic transmission or inhibitory-to-inhibitory versus inhibitory-to-excitatory transmission will disrupt persistent firing. For example, because NMDA-mediated currents are relatively stronger onto excitatory neurons than onto inhibitory neurons, disruptions in such currents break the balance between positive and negative feedback (**Fig. 4l**), with the precise size of the disruption being dependent on how asymmetrically NMDA receptors are distributed between the two pathways (**Supplementary Fig. 3** and **Supplementary Modeling**). Such relative frailty to perturbations that break the excitatory-inhibitory balance forms a prediction for the derivative-feedback models (see Discussion).

We note that the negative-derivative feedback and positive-feedback mechanisms are not mutually exclusive. Hybrid models receiving strong negative-derivative feedback and tuned positive feedback (**Fig. 4m**–**r**) could be obtained by increasing the strength of net excitatory feedback enough to offset the intrinsic decay of the neurons (**Fig. 4m**). Doing so led to networks that were both perfectly stable when properly tuned and, as a result of the strong and approximately balanced negative-derivative feedback, decayed only mildly when mistuned (**Fig. 4n**–**q**).

Irregular firing in spiking graded memory networks

A major challenge[28] to existing models of working memory has been generating the highly irregular spiking activity observed experimentally during memory periods (**Fig. 5a**). In traditional positive-feedback models, the mean synaptic input is suprathreshold and therefore drives relatively regular firing. Previous theoretical[29] and experimental[10] work instead suggests that the irregular activity seen in cortical networks results from strong inhibitory and excitatory inputs that mostly cancel on average but exhibit fluctuations that lead to a high coefficient of variation of the inter-spike intervals (CV_{isi}).

To examine irregular firing across a graded range of firing rates in the negative derivative–feedback model, we constructed a recurrently connected network of integrate-and-fire neurons consisting of excitatory and inhibitory populations with random, sparse connections between and within the populations[30]. The averaged excitation and inhibition between the populations satisfied the same balance

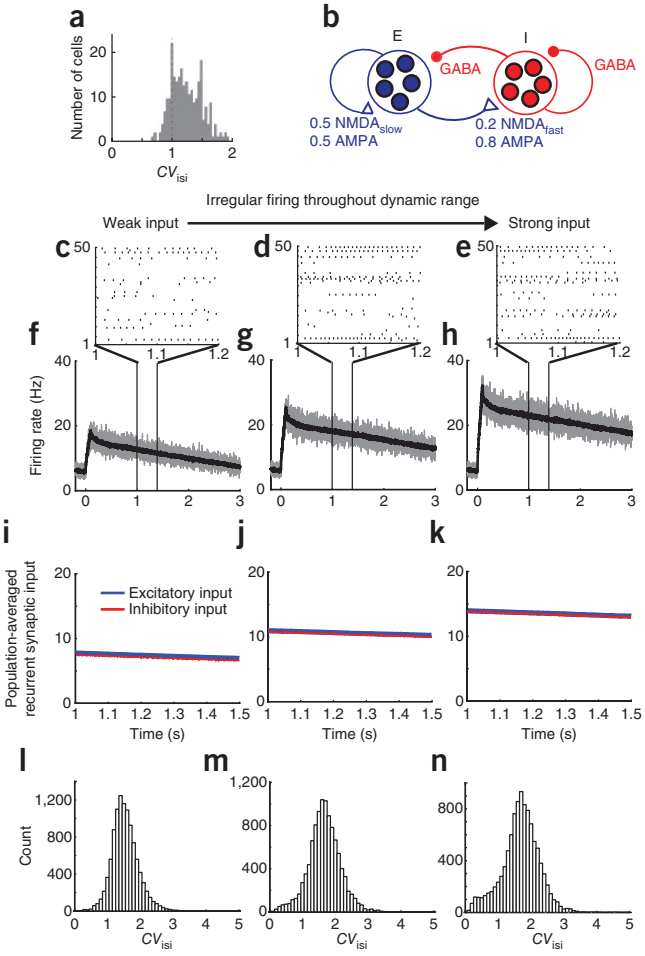

Figure 5 Irregular firing in spiking networks with graded persistent activity. (**a**) Experimentally measured irregular firing (coefficients of variation of inter-spike intervals, CV_{isi}, higher than 1) during persistent activity in a delayed-saccade task. Adapted from ref. 16. (**b**) Structure of network of spiking neurons with negative-derivative feedback. (**c**–**k**) Network response to a brief (100 ms) stimulus applied at time 0. Raster plots illustrating irregular persistent firing are shown in **c**–**e** for 50 example excitatory neurons. Instantaneous, population-averaged activity of excitatory neurons, computed in 1-ms (gray) or 10-ms (black) time bins are shown in **f**–**h**. The balance between population-averaged excitation and inhibition following offset of external input can be seen in **i**–**k**. (**l**–**n**) Histogram of CV_{isi} of active excitatory neurons during the persistent firing. Note that, for activity with strong input, a small subset of neurons fire regularly at high rate and exhibit low CV_{isi} (**n**). This reflects that the heterogeneity resulting from our simple assumption of completely randomly connected networks can result in excess positive feedback in some clusters of neurons.

condition, $J_{EE} \sim J_{EI}J_{IE}/J_{II}$, as in the firing rate models. Inhibitory currents were mediated by $GABA_A$ receptors. Recurrent excitatory currents were mediated by a mixture of AMPA and NMDA receptors (**Fig. 5b**), with a greater proportion of and slower kinetics of NMDA receptors in the excitatory feedback pathways[11–14].

As in the simpler two-population model, the network exhibited graded persistent activity whose level reflected the strength of input (**Fig. 5c**–**h**) and integrated steps of input into ramping output (**Supplementary Fig. 4**). At each maintained level, the mean synaptic inputs to each population exhibited a close balance between inhibition and excitation, with spikes triggered primarily by fluctuations away from the mean input (**Fig. 5i**–**k**). This led to the observed highly irregular activity and, as observed experimentally, a CV_{isi} distribution whose mean value exceeded 1 (**Fig. 5l**–**n**). This irregular Poisson-like firing might serve a valuable computational purpose, as Bayesian network models have suggested that Poisson firing statistics may enable probability distributions from different inputs to be combined efficiently[31,32].

Circuits with a push-pull architecture: predictions

Above, we considered a single excitatory and inhibitory population. However, neuronal recordings during parametric working memory (for example, see ref. 33) or neural integration (for example, see refs. 34,35) typically show a functional 'push-pull' organization in which competing populations of cells exhibit oppositely directed responses to a given stimulus. We found that a push-pull organization is consistent with the derivative-feedback mechanism, has additional robustness to perturbations in external inputs, and generates predictions that differentiate the derivative-feedback and traditional positive-feedback models (**Fig. 6** and **Supplementary Fig. 5**).

To construct a push-pull derivative-feedback network, we interconnected two of our two-population models (E_1 and I_1, E_2 and I_2; **Fig. 6c**) through mutual inhibitory connections (E_1 to I_2 and E_2 to I_1; **Fig. 6c**). When the circuit was tuned to have a balance of slow positive and faster negative feedback (**Supplementary Modeling**), the circuit maintained a graded range of persistent firing, with the left population increasing its firing rate when the right population decreased and vice versa (**Fig. 6f**,**l**). Persistent activity was robust to common perturbations, as in the simpler two-population models (**Fig. 4**), even when the perturbations were applied to only a single population (**Fig. 6l** and **Supplementary Fig. 5**). In addition, global shifts in background input, such as might be caused by system-wide changes in excitability, did not change the stability of persistent activity (**Supplementary Fig. 5d**), and noise caused temporally local jitter, but was largely averaged out over the long timescales of integration (**Supplementary Fig. 5h**). The former result differs from simpler models based on a single excitatory and inhibitory population, which improperly exhibit ramping activity in response to global shifts in external input; this has been suggested as a fundamental reason for the observed push-pull architectures of integrator and graded short-term memory networks[36].

A prediction for how the derivative-feedback model can be distinguished from traditional positive-feedback models is provided by examination of the intracellular currents onto the excitatory cells in each network. In the derivative-feedback models, these currents were balanced and therefore positively covaried across different levels of sustained activity (**Fig. 6i**). In contrast, in traditional positive-feedback models, inhibition was either driven by the opposing population of excitatory neurons (**Fig. 6a**) or received equal strength connections from both populations (**Fig. 6b**). In the former case, synaptic inhibition reflected the firing rates of the opposing population (**Fig. 6d**) and was anti-correlated with the excitatory inputs arriving from the same population (**Fig. 6g**). In the latter case, inhibitory neuron firing represents an average of the activity in the competing excitatory populations; if the activities of the competing excitatory populations vary symmetrically about a common background level, inhibitory neuron firing will vary only weakly with different levels of activity (**Fig. 6e**), leading to minimal correlations between inhibitory and excitatory inputs (**Fig. 6h**). If the dominant (higher firing rate) population instead varies its activity more than the non-dominant population[34], then the summed inhibition will follow the activity of the dominant population, switching when the opposite population becomes dominant and leading

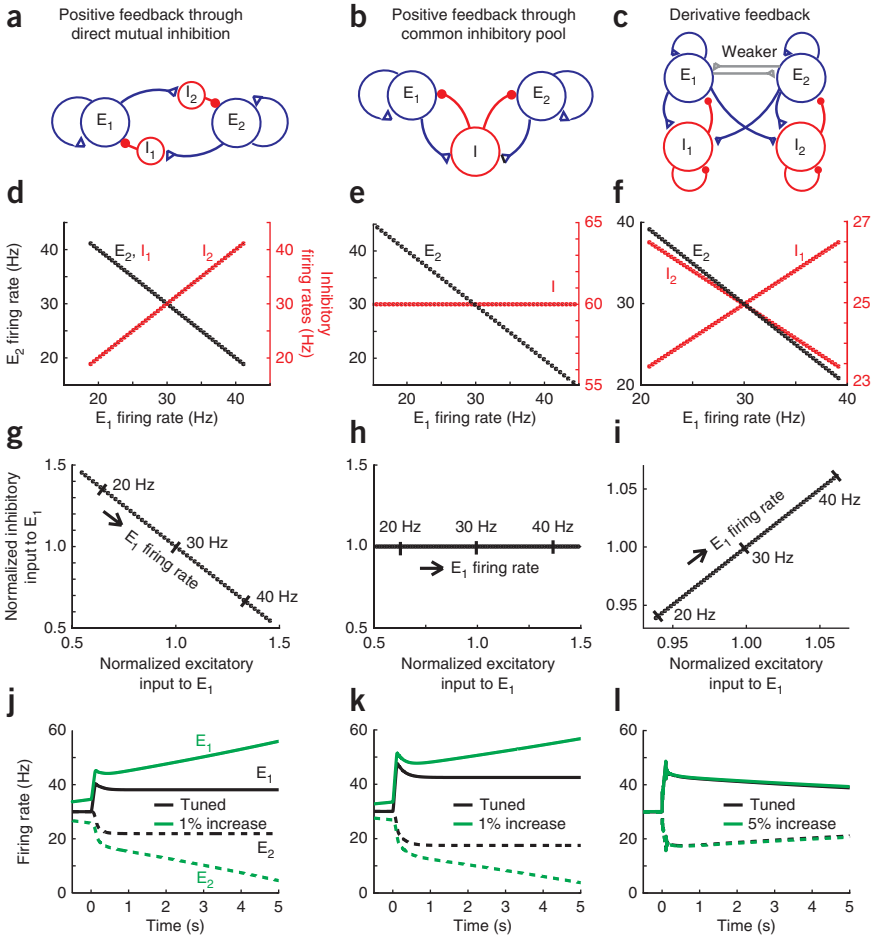

Figure 6 Synaptic inputs in derivative-feedback and common positive-feedback models. (**a**–**c**) Network structures of positive-feedback models (**a**,**b**) and derivative-feedback models (**c**) with two competing populations. (**d**–**f**) Relation between firing rates of excitatory and inhibitory neurons. Firing rates of the E_2 (black points) and inhibitory (red points) populations are plotted as a function of E_1 firing rate. (**g**–**i**) Relation between excitation and inhibition for different levels of maintained firing. *x* and *y* axes are normalized by the amount of excitation and inhibition received when the left and right excitatory populations fire at equal levels of 30 Hz. (**j**–**l**) Persistent activity in the two competing excitatory populations (solid: E_1; dashed, E_2). Perturbing the networks by uniformly increasing the intrinsic gain in E_1 leads to gross disruptions of persistent firing in positive-feedback models (green curves in **j**,**k**), but not negative derivative–feedback models (**l**). See **Supplementary Figure 5** for robustness to other perturbations.

to a non-monotonic pattern of synaptic input correlations when viewed across the entire firing rate range (data not shown).

DISCUSSION

Our results describe a mechanism for short-term memory based on negative derivative–feedback control. Networks based on this mechanism maintain activity for long durations following the offset of a stimulus and more generally act as temporal integrators of their inputs. The core requirement for negative-derivative feedback is that the pathways mediating positive and negative feedback be balanced in strength, but with slower kinetics in the positive-feedback pathways. We found that these two conditions lead to a balance between excitation and inhibition during steady persistent firing, and that this balance can be transiently disrupted by external inputs to allow a circuit to change its firing rates.

Compared with previously proposed memory networks based on positive feedback, negative derivative–feedback networks have several advantages. First, negative derivative–feedback networks inherently incorporate the observation that frontal cortical circuits have both positive- and negative-feedback pathways, with an asymmetry in the time constants of synaptic excitation onto excitatory versus inhibitory neurons[11–14]. Second, negative derivative–feedback networks are robust against many commonly studied perturbations to synaptic weights that grossly disrupt memory performance in positive-feedback models. Third, negative derivative–feedback networks inherently generate irregular firing across a graded range of persistent activity levels. These advantages are still attained in hybrid networks containing both positive- and negative-derivative feedback; thus, negative-derivative feedback is complementary to positive feedback and both mechanisms are likely to be used together in many circuits.

A balance between excitation and inhibition has been suggested as a general principle underlying the dynamics of a wide variety of cortical circuits. Physiologically, for cortical cells with large numbers of synaptic contacts and experimentally measured postsynaptic potential amplitudes, a close balance between excitation and inhibition may be essential to avoid saturation or total silencing of firing rates[37,38]. In sensory systems, the balance between inhibition and excitation includes the contribution of the external excitation driving the circuit[30,39], and activity does not persist following the offset of the stimulus. In contrast, in our study, the balance was obtained in the absence of external driving input and depended purely on recurrent synaptic inputs (or possibly a tonic background input). In bistable memory circuits, balanced excitation and inhibition[40,41] has been proposed to explain the irregular firing activity observed during elevated (UP) states of network activity[16]. However, as these models used identical time constants for the positive-feedback and negative-feedback pathways, there was no derivative feedback. As a result, they could not achieve both irregular firing activity and the graded range of persistent firing rates observed during parametric working memory and temporal integration.

A major challenge to models of graded persistent activity is maintaining the tuning of network connection strengths. In positive-feedback networks, the quantity to be tuned is the net level of network positive feedback. In negative derivative–feedback networks, the tuned quantity is the balance between excitation and inhibition. Previous foundational work in positive-feedback networks has shown that the severity of this requirement may be markedly decreased if circuit mechanisms are complemented by cellular persistence mechanisms, such as slow synaptic facilitation[25–27] or dendritic plateau potentials generated by NMDA or other voltage-activated inward currents[3,22–24,42]. Similar results hold for the derivative-feedback models if the slow process is in the excitatory-to-excitatory connections, and both dendritic plateau potentials and slow synaptic facilitation have been observed experimentally at such connections[3,26,42]. In addition, tuning

of negative-derivative feedback can be accomplished locally if neurons can monitor their balance of excitatory and inhibitory inputs. Indeed, recent experimental[43,44] and theoretical[45] work suggest that both homeostatic and developmental processes regulate this excitatory-inhibitory balance, even at the level of localized dendritic compartments[43]. The learning rules underlying the maintenance of this balance are currently unknown experimentally and are an important issue for future exploration. However, preliminary investigations suggest that a previously proposed differential Hebbian learning rule[46] may suffice to maintain the tuning of both the two-population and four-population derivative-feedback networks (**Supplementary Fig. 6**).

A separate question of robustness, focused on here, is what types of perturbations biological networks typically experience and most need to be robust against. A principle of robust control theory is that systems cannot be robust against all possible perturbations, but should be robust against common perturbations[47]. Implicitly invoking this principle, previous work has justified positive-feedback models as robust in the sense that random perturbations of connectivity only minimally affect the mean level of positive feedback[36,48], and the same argument applies to the derivative-feedback models. However, many other common perturbations, such as loss of cells or changes in neuronal gains, severely affect positive-feedback models. In contrast, derivative-feedback models can be markedly more robust to these perturbations because they produce offsetting changes in the positive and negative feedback pathways (**Fig. 4**). Derivative-feedback models are susceptible to perturbations that disrupt the excitatory-inhibitory balance of neurons, and this difference in robustness to different types of perturbations provides useful predictions. For example, we predict that completely silencing a subset of excitatory neurons would be less disruptive than silencing their synaptic inputs onto only their excitatory or only their inhibitory targets, consistent with a recent pharmacological perturbation study that showed severe disruption of persistent activity following selective targeting of NR2B-subunit containing NMDA receptors in prefrontal cortex that are primarily located at excitatory-to-excitatory synapses[14]. Similarly, we predict that globally perturbing GABAergic transmission from a subset of inhibitory neurons would be less disruptive than perturbing this input only onto its excitatory or only onto its inhibitory targets.

Slow excitation specifically in the positive-feedback pathway of negative derivative–feedback networks suggests a revision of the NMDA hypothesis for working memory storage[4,5,20,21] and deficits in schizophrenia[49]. Previously, the assumed role of NMDA receptors had been to provide a nonspecific, slow cellular time constant in all excitatory pathways[4,5,20,21]. In contrast, recent experimental studies[11–14] have reported asymmetric contributions of NMDA receptors in different feedback pathways. Building on these studies, we found an additional role of NMDA receptors in providing the delayed excitation required for negative-derivative feedback and suggest that future efforts to develop drugs for working memory disorders consider the differential contributions of NMDA receptors onto excitatory versus inhibitory target neurons.

In summary, our results describe a previously unknown mechanism for the storage of short-term memory based on corrective negative feedback. Negative feedback is a common principle of engineering control systems, in which a fundamental tenet is that strong negative feedback leads to system output (for example, an integral) that reflects the inverse of the feedback signal (for example, a derivative). Our work suggests that a similar principle is used by neocortical microcircuits for the accumulation and storage of information in working memory.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Fisher for valuable discussions and E. Aksay, K. Britten, N. Brunel, D. Butts, J. Ditterich, R. Froemke, A. Goddard, D. Kastner, B. Lankow, S. Luck, B. Mulloney, J. Raymond, J. Rinzel and M. Usrey for valuable discussions and feedback on the manuscript. We thank A. Lerchner for providing code for our initial simulations of spiking network models. This research was supported by US National Institutes of Health grants R01 MH069726 and R01 MH065034, a Sloan Foundation fellowship, and a University of California Davis Ophthalmology Research to Prevent Blindness grant.

AUTHOR CONTRIBUTIONS

S.L. and M.S.G. designed the study, analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Jonides, J. *et al.* The mind and brain of short-term memory. *Annu. Rev. Psychol.* **59**, 193–224 (2008).
2. Fuster, J.M. & Alexander, G.E. Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
3. Major, G. & Tank, D. Persistent neural activity: prevalence and mechanisms. *Curr. Opin. Neurobiol.* **14**, 675–684 (2004).
4. Durstewitz, D., Seamans, J.K. & Sejnowski, T.J. Neurocomputational models of working memory. *Nat. Neurosci.* **3**, 1184–1191 (2000).
5. Wang, X.J. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 455–463 (2001).
6. Brody, C.D., Romo, R. & Kepecs, A. Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors and dynamic representations. *Curr. Opin. Neurobiol.* **13**, 204–211 (2003).
7. Seung, H.S. How the brain keeps the eyes still. *Proc. Natl. Acad. Sci. USA* **93**, 13339–13344 (1996).
8. Machens, C.K., Romo, R. & Brody, C.D. Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science* **307**, 1121–1124 (2005).
9. Wang, X.J. Decision making in recurrent neuronal circuits. *Neuron* **60**, 215–234 (2008).
10. Haider, B. & McCormick, D.A. Rapid neocortical dynamics: cellular and network mechanisms. *Neuron* **62**, 171–189 (2009).
11. Wang, H., Stradtman, G.G., Wang, X.J. & Gao, W.J. A specialized NMDA receptor function in layer 5 recurrent microcircuitry of the adult rat prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **105**, 16791–16796 (2008).
12. Wang, H.X. & Gao, W.J. Cell type–specific development of NMDA receptors in the interneurons of rat prefrontal cortex. *Neuropsychopharmacology* **34**, 2028–2040 (2009).
13. Rotaru, D.C., Yoshino, H., Lewis, D.A., Ermentrout, G.B. & Gonzalez-Burgos, G. Glutamate receptor subtypes mediating synaptic activation of prefrontal cortex neurons: relevance for schizophrenia. *J. Neurosci.* **31**, 142–156 (2011).
14. Wang, M. *et al.* NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron* **77**, 736–749 (2013).
15. Softky, W.R. & Koch, C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci.* **13**, 334–350 (1993).
16. Compte, A. *et al.* Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. *J. Neurophysiol.* **90**, 3441–3454 (2003).
17. Haider, B., Duque, A., Hasenstaub, A.R. & McCormick, D.A. Neocortical network activity *in vivo* is generated through a dynamic balance of excitation and inhibition. *J. Neurosci.* **26**, 4535–4545 (2006).
18. Shu, Y., Hasenstaub, A. & McCormick, D.A. Turning on and off recurrent balanced cortical activity. *Nature* **423**, 288–293 (2003).
19. Murphy, B.K. & Miller, K.D. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron* **61**, 635–648 (2009).
20. Lisman, J.E., Fellous, J.M. & Wang, X.J. A role for NMDA-receptor channels in working memory. *Nat. Neurosci.* **1**, 273–275 (1998).
21. Wang, X.J. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.* **19**, 9587–9603 (1999).
22. Koulakov, A.A., Raghavachari, S., Kepecs, A. & Lisman, J.E. Model for a robust neural integrator. *Nat. Neurosci.* **5**, 775–782 (2002).
23. Goldman, M.S., Levine, J.H., Major, G., Tank, D.W. & Seung, H.S. Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. *Cereb. Cortex* **13**, 1185–1195 (2003).

24. Nikitchenko, M. & Koulakov, A. Neural integrator: a sandpile model. *Neural Comput.* **20**, 2379–2417 (2008).
25. Shen, L. Neural integration by short term potentiation. *Biol. Cybern.* **61**, 319–325 (1989).
26. Wang, Y. *et al.* Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nat. Neurosci.* **9**, 534–542 (2006).
27. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
28. Barbieri, F. & Brunel, N. Can attractor network models account for the statistics of firing during persistent activity in prefrontal cortex? *Front. Neurosci.* **2**, 114–122 (2008).
29. Vogels, T.P., Rajan, K. & Abbott, L.F. Neural network dynamics. *Annu. Rev. Neurosci.* **28**, 357–376 (2005).
30. van Vreeswijk, C. & Sompolinsky, H. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274**, 1724–1726 (1996).
31. Knill, D.C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
32. Boerlin, M. & Deneve, S. Spike-based population coding and working memory. *PLoS Comput. Biol.* **7**, e1001080 (2011).
33. Romo, R., Brody, C.D., Hernandez, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).
34. Roitman, J.D. & Shadlen, M.N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).
35. Robinson, D.A. Integrating with neurons. *Annu. Rev. Neurosci.* **12**, 33–45 (1989).
36. Cannon, S.C., Robinson, D.A. & Shamma, S. A proposed neural network for the integrator of the oculomotor system. *Biol. Cybern.* **49**, 127–136 (1983).
37. Shadlen, M.N., Britten, K.H., Newsome, W.T. & Movshon, J.A. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J. Neurosci.* **16**, 1486–1510 (1996).
38. Shadlen, M.N. & Newsome, W.T. Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol.* **4**, 569–579 (1994).
39. Destexhe, A., Rudolph, M. & Pare, D. The high-conductance state of neocortical neurons *in vivo*. *Nat. Rev. Neurosci.* **4**, 739–751 (2003).
40. Renart, A., Moreno-Bote, R., Wang, X.J. & Parga, N. Mean-driven and fluctuation-driven persistent activity in recurrent networks. *Neural Comput.* **19**, 1–46 (2007).
41. Roudi, Y. & Latham, P.E. A balanced memory network. *PLoS Comput. Biol.* **3**, 1679–1700 (2007).
42. Major, G., Polsky, A., Denk, W., Schiller, J. & Tank, D.W. Spatiotemporally graded NMDA spike/plateau potentials in basal dendrites of neocortical pyramidal neurons. *J. Neurophysiol.* **99**, 2584–2601 (2008).
43. Liu, G. Local structural balance and functional interaction of excitatory and inhibitory synapses in hippocampal dendrites. *Nat. Neurosci.* **7**, 373–379 (2004).
44. Tao, H.W. & Poo, M.M. Activity-dependent matching of excitatory and inhibitory inputs during refinement of visual receptive fields. *Neuron* **45**, 829–836 (2005).
45. Vogels, T.P., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**, 1569–1573 (2011).
46. Xie, X. & Seung, H.S. Spike-based learning rules and stabilization of persistent neural activity. in *Advances in Neural Information Processing Systems Vol. 12* (eds. Solla, S.A., Leen, T.K. & Müller, K.-R.) 199–205 (2000).
47. Csete, M.E. & Doyle, J.C. Reverse engineering of biological complexity. *Science* **295**, 1664–1669 (2002).
48. Ganguli, S. *et al.* One-dimensional dynamics of attention and decision making in LIP. *Neuron* **58**, 15–25 (2008).
49. Coyle, J.T., Tsai, G. & Goff, D. Converging evidence of NMDA receptor hypofunction in the pathophysiology of schizophrenia. *Ann. NY Acad. Sci.* **1003**, 318–327 (2003).

ONLINE METHODS

Firing rate model of one excitatory and one inhibitory population. The firing rate models of **Figure 2** were used to describe the dynamics of the average activities of, and synaptic interactions between, networks composed of one excitatory and one inhibitory population. We denote the mean firing rates of the excitatory and inhibitory populations by r_E and r_I, respectively, and the synaptic state variables for the connections from population j onto population i by s_{ij}. These firing rate and synaptic state variables are governed by the equations

$$\begin{aligned} \tau_E \dot{r}_E &= -r_E + f_E(J_{EE}s_{EE} - J_{EI}s_{EI} + J_{EO}i(t)) \\ \tau_I \dot{r}_I &= -r_I + f_I(J_{IE}s_{IE} - J_{II}s_{II} + J_{IO}i(t)) \\ \tau_{ij}\dot{s}_{ij} &= -s_{ij} + r_j \qquad \text{for } i,j = E \text{ or } I \end{aligned} \tag{5}$$

where the dot over a variable indicates differentiation with respect to time. Thus, the mean firing rate r_i approaches $f_i(x_i)$ with intrinsic time constant τ_i, where $f_i(x_i)$ represents the steady-state neuronal response to input current x_i. We consider two types of neuronal response functions: linear $f(x) = x$ (**Figs. 2c**,**d**, **3**, **4** and **6**) and a nonlinear neuronal response function (**Fig. 2c**,**d** and **Supplementary Figs. 1** and **6**) having the Naka-Rushton[50] form

$$f(x) = M\frac{(x - x_\theta)^2}{x_0{}^2 + (x - x_\theta)^2}h(x - x_\theta) \tag{6}$$

where M represents the maximal neuronal response, x_θ represents the input threshold, x_0 defines the value of $(x - x_\theta)$ at which $f(x)$ reaches its half-maximal value, and $h(x)$ denotes the step function $h(x) = 1$ for $x \geq 0$ and $h(x) = 0$ for $x < 0$.

Inputs x_i to each population include the synaptic current $J_{ij}s_{ij}$ from population j to population i and the external current $J_{iO}i(t)$, where the function $i(t)$ (not to be confused with the subscript i) denotes the temporal component of the external current. J_{ij} represents the synaptic connectivity strength onto postsynaptic neuron i from presynaptic neuron j, and the synaptic variables s_{ij} approach the presynaptic firing rate r_j with time constant τ_{ij}. We assume that one external source provides the external input to the excitatory and inhibitory populations, with J_{iO} representing the strength of the input onto population i. To model in a simple manner how stimuli are smoothed before their arrival at the memory network, we assume that the externally presented pulses of duration $t_{\text{window}} = 100$ ms (**Fig. 2c**) or step inputs (**Fig. 2d**) are exponentially filtered with time constant $t_{\text{ext}} = 100$ ms.

In **Figure 2b**, we performed a firing-rate clamp experiment to illustrate how recurrent excitatory and inhibitory inputs provide negative derivative–like feedback in response to a change in firing rate. In this experiment, in which r_E steps between two fixed levels, the external input to the excitatory population in equation (5) is adjusted so that the profile of r_E becomes a step function $h(t)$. The remaining variables then are allowed to vary following the equations given in equation (5).

In **Figures 3** and **4**, we consider networks with a mixture of two different types of synapses, NMDA type and AMPA type, in both of the excitatory pathways (from excitatory to excitatory and excitatory to inhibitory). Thus, the excitatory and inhibitory populations receive both types of excitatory synaptic inputs and the model is given by

$$\begin{aligned} \tau_i \dot{r}_i &= -r_i + f_i\left(J^N_{iE}s^N_{iE} + J^A_{iE}s^A_{iE} - J_{iI}s_{iI} + J_{iO}i(t)\right) \\ \tau^k_{ij}\dot{s}^k_{ij} &= -s^k_{ij} + r_j \text{ where } i,j = E \text{ or } I, \text{ and } k = N \text{ or } A \end{aligned} \tag{7}$$

The superscripts N and A denote NMDA-type and AMPA-type synapses, respectively, and all other variables are the same as in equation (5). In **Figure 3a**, the strengths of total excitatory synaptic currents and the fractions of NMDA-type synapses are represented by J_{iE} and q_{iE}; that is, $J_{iE} = J^N_{iE} + J^A_{iE}$ and $q_{iE} = J^N_{iE}/J_{iE}$ for $i = E$ or I. In the purely negative derivative–feedback models of **Figure 4g–l**, the network connectivity is tuned to have no net positive feedback by setting the strengths of positive and negative feedback to be precisely equal through the relation $J_{EE} = J_{EI}J_{IE}/(1 + J_{II})$. On the other hand, in the hybrid models of **Figure 4m–r**, excess positive feedback is tuned to precisely cancel the leakage by setting $J_{EE} - J_{EI}J_{IE}/(1 + J_{II}) = 1$.

Throughout our study, we set the intrinsic time constants of excitatory and inhibitory neurons, τ_E and τ_I, to 20 ms and 10 ms, respectively[51]. The time constants of $GABA_A$-type inhibitory synapses, τ_{EI} and τ_{II}, were set to 10 ms (refs. 52,53). Based on experimental measurements of excitatory synaptic currents in prefrontal cortex[13], the time constants of excitatory synaptic currents and the fractions of NMDA-mediated synaptic currents are set as follows: in the networks with a mixture of NMDA- and AMPA-mediated excitatory currents (**Figs. 3** and **4**), $\tau^N_{EE} = 150$ ms and $\tau^A_{EE} = 50$ ms in excitatory neurons, and $\tau^N_{IE} = 45$ ms and $\tau^A_{IE} = 20$ ms in inhibitory neurons. Note that these time constants reflect the kinetics of postsynaptic potentials observed to be triggered by activation of NMDA- or AMPA-type receptors, and likely include the effects of additional intrinsic ionic conductances, as these experiments were performed without blocking intrinsic ionic currents[13]. The fractions of NMDA-mediated synaptic currents in excitatory neurons and inhibitory neurons, q_{EE} and q_{IE}, were set to 0.5 and 0.2, respectively. The time constants of excitatory synapses for networks with only a single type of synaptic current for each connection in **Figure 2** were set to $\tau_{EE} = 100$ ms and $\tau_{IE} = 25$ ms to satisfy the average excitatory kinetics $\tau_{EE} = q_{EE}\tau^N_{EE} + (1 - q_{EE})\tau^A_{EE}$ and $\tau_{IE} = q_{IE}\tau^N_{IE} + (1 - q_{IE})\tau^A_{IE}$. Note that, because $\tau_{EE} > \tau_{IE}$, this provides slower positive than negative feedback (see equation (4)). The synaptic strengths J_{ij} were set to satisfy the balance conditions given by equation (3) (**Supplementary Modeling**).

We note that our model can similarly be extended to include both fast ($GABA_A$) and slow ($GABA_B$) components of synaptic transmission. In this case, the conditions for negative-derivative feedback have the same form as considered previously, but with replacement of τ_{II} and τ_{EI} by $\tau_{II} = q_{II}\tau^{GB}_{II} + (1 - q_{II})\tau^{GA}_{II}$ and $\tau_{EI} = q_{EI}\tau^{GB}_{EI} + (1 - q_{EI})\tau^{GA}_{EI}$, where the superscripts GA and GB denote the fast ($GABA_A$) and slow ($GABA_B$) components and q_{EI} and q_{II} denote the proportion of $GABA_B$ currents. **Supplementary Figure 7** shows an example simulation with inclusion of such a slow, inhibitory component of synaptic transmission.

Firing rate model of two competing populations. In **Figure 6**, we compared networks of competing populations using positive-feedback control versus negative derivative–feedback control. The connectivity between populations varies in different models but the dynamics of the firing rates and the synapses are the same as in equation (5)

$$\begin{aligned} \tau_i \dot{r}_i &= -r_i + f_i\left(\sum_j J_{ij}s_{ij} + J_{iO}i(t) + J_{i,\text{tonic}}\right) \\ \tau_{ij}\dot{s}_{ij} &= -s_{ij} + r_j \quad \text{where } i,j = E_1, I_1, E_2, \text{ or } I_2 \end{aligned} \tag{8}$$

Here, E and I represent excitatory and inhibitory populations, respectively, and the subscripts 1 or 2 are the index of the population. The temporal component of $i(t)$ is the same transient pulse-like input as in the firing rate model of equation (5) and $J_{i,\text{tonic}}$ is the strength of the tonic input.

In the positive-feedback network with direct mutual inhibition (**Fig. 6a**,**d**,**g**,**j**), population E_i receives recurrent excitatory input $J_{E_iE_i}s_{E_iE_i}$ and inhibitory input $J_{E_iI_i}s_{E_iI_i}$ from the same population, and external inputs $J_{E_iO}i(t)$ and $J_{E_i,\text{tonic}}$. The inhibitory subpopulation I_i, for $i = 1$ or 2, receives only the excitatory inputs $J_{I_iE_j}s_{I_iE_j}$ from the opposing population ($j = 2$ or 1, respectively).

The positive-feedback network with a common inhibitory pool (**Fig. 6b**,**e**,**h**,**k**) is composed of three populations: two excitatory populations E_1 and E_2, and the common inhibitory population I. E_i receives recurrent excitatory input $J_{E_iE_i}s_{E_iE_i}$ from itself, inhibitory input $J_{E_iI}s_{E_iI}$, and external inputs $J_{E_iO}i(t)$ and $J_{E_i,\text{tonic}}$. The common inhibitory population I receives input $J_{IE_1}s_{IE_1}$ from E_1 and input $J_{IE_2}s_{IE_2}$ from E_2.

In the negative derivative-feedback model with two competing populations (**Figs. 6c**,**f**,**i**,**l** and **Supplementary Figs. 5** and **6e–h**), each population has the same structure as in the single population in equation (5). Connections between the two competing populations are mediated by projections from the excitatory cells of each population that project weakly onto excitatory cells of the opposing population and more strongly onto inhibitory cells of the opposing population. Thus, the excitatory subpopulation E_i receives inputs $J_{E_iE_i}s_{E_iE_i}$ and $J_{E_iI_i}s_{E_iI_i}$ from the same side, $J_{E_iE_j}s_{E_iE_j}$ from the opposite side, and external inputs $J_{E_iO}i(t)$

and $J_{E_i,\text{tonic}}$. Similarly, the inhibitory subpopulation I_i receives inputs $J_{I_iE_i}s_{I_iE_i}$ and $J_{I_iI_i}s_{I_iI_i}$ from the same side, and $J_{I_iE_j}s_{I_iE_j}$ from the opposite side.

The intrinsic time constants of excitatory and inhibitory neurons and the synaptic time constants are the same as in the single population (the remaining parameters are given in **Supplementary Modeling**). All the simulations of the firing rate models were run with a fourth-order explicit Runge-Kutta method using the function ode45 in MATLAB.

Spiking network model with leaky integrate-and-fire neurons. In **Figure 5** and **Supplementary Figure 4**, we constructed a recurrent network of excitatory and inhibitory populations of spiking neurons with balanced excitation and inhibition. We found that this spiking network maintained graded levels of persistent activity with temporally irregular firing. Here, we describe the dynamics of individual neuron activity and the synaptic currents connecting the neurons.

The spiking network consists of N_E excitatory and N_I inhibitory current-based leaky integrate-and-fire neurons that emit a spike when a threshold is reached and then return to a reset potential after a refractory period. These neurons are recurrently connected to each other and receive transient stimuli from an external population of N_O neurons (**Fig. 5b**, external population not shown). The connectivity between neurons is sparse and random with connection probability p so that, on average, each neuron receives N_Ep, N_Ip and N_Op synaptic inputs from the excitatory, inhibitory, and external populations, respectively.

The dynamics of the subthreshold membrane potential V of the lth neuron in population i, and the dynamics of the synaptic variable $s_{ij}^{lm,k}$ onto this neuron from the mth neuron in population j are

$$\tau_i \frac{dV_i^l}{dt} = -(V_i^l - V_L) + \sum_m \tilde{J}_{iE} p_{iE}^{lm} (q_{iE}^N s_{iE}^{lm,N}(t) + q_{iE}^A s_{iE}^{lm,A}(t)) - \sum_m \tilde{J}_{iI} p_{iI}^{lm} s_{iI}^{lm}(t) + \sum_m \tilde{J}_{iO} p_{iO}^{lm} s_{iO}^{lm}(t) \quad (9)$$

$$\tau_{ij}^k \frac{ds_{ij}^{lm,k}}{dt} = -s_{ij}^{lm,k} + \sum_{t_j^m} \delta(t - t_j^m), \text{ for } j = E, I, \text{ or } O \text{ and } k = N \text{ or } A \quad (10)$$

The first term on the right-hand side of equation (9) corresponds to a neuronal intrinsic leak process such that, without the input, the voltage decays to the resting potential V_L with time constant τ_i. The second term is the sum of the recurrent NMDA- and AMPA-mediated excitatory synaptic currents as in equation (7). The dynamic variables $s_{iE}^{lm,N}$ and $s_{iE}^{lm,A}$ represent NMDA- and AMPA-mediated synaptic currents from cell m of the excitatory population. The sum of the strengths of NMDA- and AMPA-mediated synaptic currents, and the fractions of NMDA- and AMPA-mediated currents, are assumed to be uniform across the population and are denoted by $\tilde{J}_{iE}$, q_{iE}^N and $q_{iE}^A = 1 - q_{iE}^N$, respectively. p_{iE}^{lm} is a binary random variable with probability p representing the random connectivity between neurons. Similarly, the third and fourth terms represent the total synaptic inputs from the inhibitory population and the external population. As for the excitatory currents, the dynamic variables s_{iI}^{lm} and s_{iO}^{lm} denote the synaptic currents with strengths $\tilde{J}_{iI}$ and $\tilde{J}_{iO}$, respectively, and p_{iI}^{lm} and p_{iO}^{lm} are binary random variables with probability p.

In the dynamics of $s_{ij}^{lm,k}$ in equation (10), a presynaptic spike at time t_j^m from neuron m in population j causes a discrete jump in synaptic current followed by an exponential decay with time constant τ_{ij}^k. Here, the spikes in the external population, representing inputs to be remembered, are generated by a Poisson process with rate r_O during a time window t_{window} (**Fig. 5**, with $r_O = 0$ during the memory period) or with rate r_O after $t = 0$ (**Supplementary Fig. 4**). Note that the strength of $s_{ij}^{lm,j}$, denoted by $\tilde{J}_{ij}$ in equation (9), corresponds to the integrated area under a single postsynaptic potential, and not the height of a single postsynaptic potential. Furthermore, the connectivity strengths $\tilde{J}_{ij}$ were scaled as

$$\tilde{J}_{ij} = \hat{J}_{ij} / \sqrt{N_j p} \quad \text{for fixed } \hat{J}_{ij} \quad (11)$$

This scaling made the fluctuations in the input remain of the same order of magnitude as the mean input as the network size varied[30].

In **Figure 5l–n**, the coefficients of variation of the inter-spike intervals were computed for 3 s from time 300 ms to 3300 ms using all excitatory neurons that exhibited more than 5 spikes during this period.

In the simulation, $N_E = 16{,}000$, $N_I = 4{,}000$, $N_O = 20{,}000$ and $p = 0.1$. The time constants and the fractions of NMDA-mediated currents were the same as in the firing rate models: $\tau_E = 20$ ms, $\tau_I = 10$ ms, $\tau_{EI} = \tau_{II} = 10$ ms, $\tau_{EE}^N = 150$ ms, $\tau_{EE}^A = 50$ ms, $\tau_{IE}^N = 45$ ms, $\tau_{IE}^A = 20$ ms, $q_{EE}^N = 0.5$ and $q_{IE}^N = 0.2$. The parameters for the synaptic strengths were tuned to achieve a balance between excitatory and inhibitory inputs during sustained activity (remaining parameters are given in **Supplementary Modeling**).

The numerical integration of the network simulations was performed using the second-order Runge-Kutta algorithm. Spike times were approximated by linear interpolation, which maintains the second-order nature of the algorithm[54].

50. Wilson, H.R. *Spikes, Decisions and Actions* (Oxford University Press, 1999).
51. McCormick, D.A., Connors, B.W., Lighthall, J.W. & Prince, D.A. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *J. Neurophysiol.* **54**, 782–806 (1985).
52. Salin, P.A. & Prince, D.A. Spontaneous $GABA_A$ receptor–mediated inhibitory currents in adult rat somatosensory cortex. *J. Neurophysiol.* **75**, 1573–1588 (1996).
53. Xiang, Z., Huguenard, J.R. & Prince, D.A. GABAA receptor-mediated currents in interneurons and pyramidal cells of rat visual cortex. *J. Physiol.* (*Lond.*) **506**, 715–730 (1998).
54. Hansel, D., Mato, G., Meunier, C. & Neltner, L. On numerical simulations of integrate-and-fire neural networks. *Neural Comput.* **10**, 467–483 (1998).