

A Technical Appendices and Supplementary Material

The code and data to reproduce the results in this paper can be found on Github. Spiking network simulations were performed on a single CPU using the Aurn simulator [34]. The perturbations of rule quadruplets in fig. 5C (≈ 3000 simulations) were performed on 500 CPUs on the ISTA HPC cluster. Rate models and analysis of all simulations was performed using numpy [35], JAX [36], matplotlib [37], SciPy [38] and scikit-learn [39].

A.1 Recurrent spiking network model

A.1.1 Network model

We considered a recurrent spiking network with $N_E = 4096$ excitatory neurons and $N_I = 1024$ inhibitory neurons (leaky-integrate and fire point neurons with variable threshold, AMPA and NMDA currents, and conductance-based synapses). This network was based on previous work [8, 20]. The membrane potential dynamics of neuron j (excitatory or inhibitory) followed:

$$\tau_m \frac{d}{dt} V_j(t) = - (V_j(t) - V_{\text{rest}}) - g_j^E(t) (V_j(t) - E_E) - g_j^I(t) (V_j(t) - E_I), \quad (9)$$

with $\tau_m = 20$ ms, $V_{\text{rest}} = -70$ mV, $E_E = 0$ mV and $E_I = -80$ mV.

A postsynaptic spike occurred whenever the membrane potential $V_j(t)$ crossed a threshold $V_j^{\text{th}}(t)$, with an instantaneous reset to $V_{\text{reset}} = -70$ mV. This threshold $V_j^{\text{th}}(t)$ was incremented by $V_{\text{spike}}^{\text{th}} = 100$ mV every time neuron j spiked and otherwise decayed following:

$$\tau_{\text{th}} \frac{d}{dt} V_j^{\text{th}}(t) = V_{\text{base}}^{\text{th}} - V_j^{\text{th}}(t), \quad (10)$$

with $V_{\text{base}}^{\text{th}} = -50$ mV. The excitatory and inhibitory conductances, g^E and g^I evolved such that

$$\begin{aligned} g_j^E(t) &= a g_j^{\text{AMPA}}(t) + (1 - a) g_j^{\text{NMDA}}(t) \quad \text{and} \\ \frac{d}{dt} g_j^I(t) &= -\frac{g_j^I(t)}{\tau_{\text{GABA}}} + \sum_{i \in \text{Inh}} w_{ij}(t) S_i(t) \\ \text{with } \frac{d}{dt} g_j^{\text{AMPA}}(t) &= -\frac{g_j^{\text{AMPA}}(t)}{\tau_{\text{AMPA}}} + \sum_{i \in \text{Exc}} w_{ij}(t) S_i(t) \quad \text{and} \\ \frac{d}{dt} g_j^{\text{NMDA}}(t) &= \frac{g_j^{\text{AMPA}}(t) - g_j^{\text{NMDA}}(t)}{\tau_{\text{NMDA}}}, \end{aligned} \quad (11)$$

with $w_{ij}(t)$ the connection strength between neurons i and j (unitless), $a = 0.23$ (unitless), $\tau_{\text{GABA}} = 10$ ms, $\tau_{\text{AMPA}} = 5$ ms, $\tau_{\text{NMDA}} = 100$ ms, $S_i(t) = \sum \delta(t - t_i^*)$ the spike train of presynaptic neuron i , where t_i^* denotes the spike times of neuron k , and δ the Dirac delta.

The network was initialized with random sparse connectivity (10%), with $w_{EE}^{\text{init}} = w_{EI}^{\text{init}} = 0.1$ and $w_{IE}^{\text{init}} = w_{II}^{\text{init}} = 1$.

The excitatory neurons in the network received $N_{\text{inp} \rightarrow E} = 11025$ inputs from Poisson neurons firing at $r_{\text{bg}}^{\text{inp}} = 10$ Hz. When a stimulus was active, a subset of the input neurons increased their firing rate to $r_{\text{active}}^{\text{seq}} = 100$ Hz. The connectivity from input neurons to excitatory and inhibitory neurons was receptive-field-like: for each recurrent neuron, we selected a random input neuron as the center of the circular receptive field of radius 8. The connections from neurons of this circular patch of input neurons to the considered recurrent neuron was $w_{\text{inp}} = 0.075$, and 0 to all other input neurons. The inhibitory neurons received inputs from $N_{\text{inp} \rightarrow I} = 4096$ Poisson neurons with w_{inp} and similar receptive field connectivity than for the excitatory population. However, inhibitory neurons only received background inputs ($r_{\text{bg}}^{\text{input}}$) and no specific stimulus patterns.

A.1.2 Plasticity parameterization

This parameterization of plasticity rules included variations of spike-timing-dependent plasticity, and was taken from previous work [19, 20]. The weight from neuron i to neuron j of type X and Y (excitatory or inhibitory) evolved such that:

$$\frac{dw_{ij}(t)}{dt} = \eta[S_i(t)(\alpha_{XY} + \kappa_{XY}x_j(t)) + S_j(t)(\beta_{XY} + \gamma_{XY}x_i(t))] \quad (12)$$

with $\eta = 0.01$ a fixed learning rate, $S_i(t) = \sum_k \delta(t - t_k^i)$ the spike train of neuron i , δ the Dirac delta function to denote the presence of a pre (post)-synaptic spike at time t . The synaptic traces x_i and x_j are low-pass filters of the activity of presynaptic neuron i and postsynaptic neuron j , with time constants τ_{pre} and τ_{post} , such that:

$$\frac{d}{dt}x_i(t) = -\frac{x_i(t)}{\tau_{\text{pre}}^{XY}} + S_i(t) \quad \text{and} \quad \frac{d}{dt}x_j(t) = -\frac{x_j(t)}{\tau_{\text{post}}^{XY}} + S_j(t), \quad (13)$$

Overall, this search space comprised 6 tunable plasticity parameters per synapse type XY ($X, Y \in \{E, I\}$): $\theta_{XY} = [\alpha_{XY}, \beta_{XY}, \gamma_{XY}, \kappa_{XY}, \tau_{\text{pre}}^{XY}, \tau_{\text{post}}^{XY}]$, for a total of 24 plasticity parameters across all four synapse types.

Note that all weights in the network were capped at all times, in the $[0, w_{\text{max}}]$ range, with $w_{\text{max}} = 20$, though rule quadruplets in [20] were considered unstable if more than 10% of the weights at any synapse type reached these extreme values.

A.1.3 Familiarity detection task

We considered a familiarity detection task that was similar to previous work [20]. The network first received nonspecific background inputs for 1h—pre-training phase—, followed by a 40s training phase during which four non-overlapping input patterns—the familiar stimuli—were active in alternation. When a given stimulus was active, $\approx 10\%$ of the input neurons to the excitatory population had elevated firing rates, while the others remained at background. After training, we reverted to background inputs and regularly probed the network with familiar and novel stimuli for an hour—post-training phase.

Given the input structure and connectivity described above, each stimulus, novel or familiar excited a different subset of recurrent neurons. We defined “engrams” for each stimulus pattern (novel or familiar) by probing the network before training started, and labeling the top 10% of excitatory and inhibitory neurons as part of the engram for the presented stimulus. In practice, since the stimuli were non overlapping, the engrams defined this way also had little overlap.

Note that each stimulus elicited a different network response, in the sense that each stimulus preferentially excited a different subset of recurrent neurons (this can be seen in fig. S2) and thus the stimulus identity could be decoded from the population vector of neuron activities. However, whether this stimulus has been encountered by the network in the past—its novelty or familiarity— was unknown.

We chose a simple decoding strategy for stimulus familiarity: the mean firing rate of the excitatory population (each excitatory neuron contributes equally to the decoding). In the paper that inspired this work [20], a Student t-test was performed over several seeds/simulations/stimuli over mean firing rates in response to familiar or novel stimuli to determine whether novel stimuli elicited statistically different mean firing rates than familiar stimuli. Note that by design, all stimuli elicited statistically indistinguishable mean firing rates in static or naive plastic networks. Thus this task flagged a stimulus-specific change in network activity due to synaptic plasticity.

A.2 Feedforward spiking network model

A.2.1 Network model

1000 Poisson neurons (800 excitatory and 200 inhibitory) projected onto a single output neuron, with the same neuron model and parameters as in the previous section. The excitatory weights were fixed at $w_{ee} = 0.1$, the inhibitory weights were plastic with the parameterization defined for the recurrent spiking case, and initialized at $w_{ie} = 1$.

Note that in fig. 5C, we used a different learning rule, not part of the plasticity parameterization described above. For this rule, $\frac{dw}{dt} = \beta S_{\text{pre}}(t) + C$ with $\beta = 0.4$ and $C = 0.00012$ to obtain a target firing rate of 3Hz given the integration time-step of the simulation (0.1ms).

A.2.2 Familiarity detection task

The task closely resembled the task in the recurrent case. During a pre-training phase of 1h, all input neurons fired at 10Hz. During the training phase that lasted 60s, 100 excitatory and 25 inhibitory increased their firing rates to 100Hz and 50Hz respectively (the "familiar" stimulus). After the training phase, the network was regularly probed on its network response to the familiar stimulus and another, novel stimulus of the same structure than the familiar stimulus but with different neurons (no overlap).

A.3 Toy model 1: Explicit parameterization

A.3.1 2D version

The model is a linear feedforward network with two inputs $\mathbf{x} = (x_0, x_1)$ projecting on a single output neuron y :

$$y(t) = w_0(t)x_0(t) + w_1(t)x_1(t) \quad (14)$$

Besides, $\forall t \geq 0$, $y(t) \geq 0$, $x_0(t) \geq 0$, $x_1(t) \geq 0$; $\|\mathbf{x}\| = 1$ with $\|\cdot\|$ the L2 norm. Weights were plastic and unconstrained.

Initially, we considered a four-parameter set of Hebbian/non-Hebbian plasticity rules inspired by the full spiking model (see mean-field section below for the relationship between spike-based and rate-based plasticity):

$$\frac{\partial w_i(t)}{\partial t} = \eta(\theta_0 + \theta_1 x_i(t) + \theta_2 y(t) + \theta_3 x_i(t)y(t)) \quad (15)$$

with $\theta_0, \theta_1, \theta_2$ and θ_3 four plasticity parameters, and $\eta = 0.01$ a fixed learning rate (omitted below). From this full search space, we only considered rules that admitted a target output firing rate $y^* \in \mathbb{R}^+$ as a stable fixed point for all inputs \mathbf{x} considered here. The existence of y^* as a fixed point implied that for all inputs x_i

$$\theta_0 + \theta_1 x_i + \theta_2 y^* + \theta_3 x_i y^* = 0 \implies \theta_0 = -\theta_2 y^* \text{ and } \theta_1 = -\theta_3 y^* \quad (16)$$

Thus the search space became two-dimensional:

$$\begin{cases} \frac{\partial w_0}{\partial t} = (y - y^*)(\theta_0 + \theta_1 x_0) \\ \frac{\partial w_1}{\partial t} = (y - y^*)(\theta_0 + \theta_1 x_1) \end{cases} \implies \dot{\mathbf{w}} = A\mathbf{w} + B \quad (17)$$

with $A = \begin{pmatrix} x_0(\theta_0 + \theta_1 x_0) & x_1(\theta_0 + \theta_1 x_0) \\ x_0(\theta_0 + \theta_1 x_1) & x_1(\theta_0 + \theta_1 x_1) \end{pmatrix}$ and $B = -y^* \begin{pmatrix} \theta_0 + \theta_1 x_0 \\ \theta_0 + \theta_1 x_1 \end{pmatrix}$.

This system has only one non-zero eigenvalue: $\lambda_1 = \theta_0(x_0 + x_1) + \theta_1(x_0^2 + x_1^2)$. The system thus has a neutral mode ($\lambda_0 = 0$), for the system to converge to a point on the line attractor, we need $\lambda_1 < 0$. Because $\|\mathbf{x}\| = 1$ and $x_0, x_1 \geq 0$, $x_0 + x_1 \in [1, \sqrt{2}]$ and $x_0^2 + x_1^2 = 1$. As a result, we need θ_0 and θ_1 to be below the lines $\theta_0 + \theta_1 = 0$ and $\theta_0\sqrt{2} + \theta_1 = 0$.

For the numerical results reported in the paper, we simulated the system in the A-B-A task with: $y^* = 1$, $\mathbf{w}_0 = (1, 0.27) \in \mathbf{W}_{\mathbf{x}_{\text{bg}}}^*$, $\mathbf{x}_{\text{bg}} \angle \mathbf{x}_{\text{stim}} = \frac{\pi}{4}$. We ran the system until convergence at each phase of the A-B-A task, in practice we found that $T = 20000$ epochs was sufficient.

We verified that the results of the parameter sweeps on θ_0 and θ_1 had similar trends for different stimuli angles and initializations.

A.3.2 Extension to higher input dimensions

In the main text, we chose the smallest model possible (2D) for ease of visualization. However, the findings readily extend to higher dimensions ($N_{\text{in}} > 2$ input neurons).

We consider N_{in} input neurons with activity \mathbf{x} projecting on a single output neuron y with weights \mathbf{w} : $y = \mathbf{W}^T \mathbf{x}$. We choose \mathbf{x} to be of unit norm with nonnegative entries. The plasticity rules are the same as in the 2D case:

$$\frac{\partial w_i}{\partial t}(t) = (y(t) - y^*)(\theta_0 + \theta_1 x_i(t)), \quad i = 1, \dots, N_{\text{in}} \quad (18)$$

This is an affine system of ODEs, which can be written in vector form as $\dot{\mathbf{w}} = A\mathbf{w} + \mathbf{b}$, with $A = \mathbf{x}(\theta_0 \mathbf{1} + \theta_1 \mathbf{x})$, and $\mathbf{1}$ a N_{in} -dimensional vector of ones.

For a constant input \mathbf{x} , this system is rank one, and the non-zero eigenvalue is $\lambda = \theta_0 \sum_{i=1}^{N_{\text{in}}} x_i + \theta_1 \sum_{i=1}^{N_{\text{in}}} x_i^2 = \theta_0 \sum_{i=1}^{N_{\text{in}}} x_i + \theta_1$ for unitary norm inputs. Note that this is a generalization of the derivation presented above.

Edge of stability: For the system to be stable, we need $\lambda < 0$, which translate for unit-norm, nonnegative inputs to $\theta\sqrt{N_{\text{in}}} + \theta_1 < 0$ and $\theta_0 + \theta_1 < 0$.

Defining a metric to evaluate memory: As in the 2D case, we define \mathbf{w}_{bg} , the steady state weights for input \mathbf{x}_{bg} at the start of the A-B-A task, and $\mathbf{w}_{\text{bg}'}$ are those for \mathbf{x}_{bg} at the end of the task. However, the 2D definition of RI (eq. (4)) does not generalize readily to N-dimensions, as $\mathbf{w}_{\text{bg} \cap \text{stim}}$, the intersection between the hyperplanes $\mathbf{W}_{\mathbf{x}_{\text{bg}}}^* : \mathbf{w}^T \mathbf{x}_{\text{bg}} = y^*$ and $\mathbf{W}_{\mathbf{x}_{\text{stim}}}^* : \mathbf{w}^T \mathbf{x}_{\text{stim}} = y^*$ is not unique (assuming these hyperplanes are not parallel). As a proxy, we define $RI = \|\mathbf{w}_{\text{bg}'} - \mathbf{w}_{\text{bg}}\|$, which only evaluates how far the final state is from the initial one, and not whether the change is pushing towards the intersection or not.

Overall, as can be seen in fig. S8 increasing the dimensionality of the toy models did not change qualitatively the findings reported in the main paper.

A.4 Toy model 2: Implicit parameterization

A.4.1 2D version

This toy model only described the fixed point reached by an implicitly-defined class of learning rules operating in the same linear feedforward network as in previous section.

Specifically, we made two assumptions on the learning rules:

$$\text{Stabilization: } \forall (\mathbf{x}, y_0, \mathbf{w}_0) \in \mathbb{R}^{2 \times 1 \times 2}, \quad \lim_{t \rightarrow +\infty} y(t, \mathbf{x}, y_0, \mathbf{w}_0) = y^* \quad (19)$$

$$\text{Distance minimization: } \forall (\mathbf{x}, y_0, \mathbf{w}_0) \in \mathbb{R}^{2 \times 1 \times 2}, \quad \lim_{t \rightarrow +\infty} \mathbf{w}(t, \mathbf{x}, y_0, \mathbf{w}_0) = \underset{\mathbf{w} \in \mathbf{W}_{\mathbf{x}}^*}{\operatorname{argmin}} \|\mathbf{w} - \mathbf{w}_0\|_{\Sigma}^2 \quad (20)$$

with $\|\cdot\|_{\Sigma}$ the norm induced by the Mahalanobis distance D :

$$\forall (\mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^{2 \times 2}, \quad D_{\Sigma}(\mathbf{w}_1, \mathbf{w}_2) = \sqrt{(\mathbf{w}_1 - \mathbf{w}_2)^T \Sigma^{-1} (\mathbf{w}_1 - \mathbf{w}_2)} \quad (21)$$

with $\Sigma^{-1} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, where a, b , and c are the three plasticity parameters in this search space. To be a well-defined distance, Σ^{-1} needs to be positive semi-definite leading to the further constraint that $a \geq 0$ and $ac - b^2 \geq 0$.

From these assumptions, we derived the final network state $\mathbf{w}^f = (w_0^f, w_1^f)$ as a function of the initialization $\mathbf{w}^i = (w_0^i, w_1^i)$, \mathbf{x}^i (input for which the network is initialized), and (fixed) input \mathbf{x}^f . Since the final state belongs to $\mathbf{W}_{\mathbf{x}^f}^*$, we have $\mathbf{w}^{fT} \mathbf{x}^f = y^* \implies w_1^f = \frac{y^* - w_0^f x_0^f}{x_1^f}$. We minimize the distance D :

$$D(\mathbf{w}, \mathbf{w}^i)^2 = \frac{\|\mathbf{x}^f\|_{\Sigma^{-1}}^2 - 1}{x_1^f{}^2} w_0^2 + 2 \frac{-aw_0^i x_1^f{}^2 + bx_1^f [y^* + w_0^i x_0^f - w_1^i x_1^f] + cx_0^f [w_1^i x_1^f - y^*]}{x_1^f{}^2} w_0 \quad (22)$$

$$+ \|\mathbf{w}^i\|_{\Sigma}^2 + \frac{y^* [-2bw_0^i x_1^f + c(-2w_1^i x_1^f + y^*)]}{x_1^f{}^2} \quad (23)$$

Since $\frac{\|\mathbf{x}^f\|_{\Sigma^{-1}}^2 - 1}{x_1^f{}^2} > 0$, the expression above has a single minimum:

$$w_0^f = \frac{aw_0^i x_1^f{}^2 + bx_1^f (w_1^i x_1^f - w_0^i x_0^f - y^*) + cx_0^f (y^* - w_1^i x_1^f)}{\|\mathbf{x}^f\|_{\Sigma^{-1}}^2 - 1} \quad (24)$$

We applied the result above twice, for each phase of the A-B-A task.

Relationship between explicit and implicit toy models: although we don't have a general mapping from the implicit to the explicit rules in both toy models, some rules have the same steady states in both cases.

Notably, the Hebbian rule $\frac{\partial w_i}{\partial t} = -x_i(y - y^*)$ ($\theta_0 = 0, \theta_1 = -1$) had identical fixed point to the rule minimizing the Euclidean distance in the implicit model ($a = 1, b = 0, c = 1$), see fig. S6D. The explicit form of this rule could also be seen as the gradient descent update wrt the loss function $\mathcal{L} = \frac{(y - y^*)^2}{2}$, i.e. $\frac{\partial \mathcal{L}}{\partial w_i} = x_i(y - y^*) = -\frac{\partial w_i}{\partial t}$.

A.4.2 Extension to higher input dimensions

This toy model has the same network architecture as above, but the number of input neurons does change the number of plasticity parameters, as the rules are this time parameterized by a distance metric of the Mahalanobis family (with covariance $\Sigma \in \mathbb{R}^{N_{in} \times N_{in}}$, leaving us with $\frac{N_{in}(N_{in}+1)}{2}$ plasticity parameters.

Edge of stability: This corresponds to Σ losing its positive semi-definite property (i.e. at least one eigenvalue becomes 0).

A.5 Toy-model 3: Mean-field-inspired model

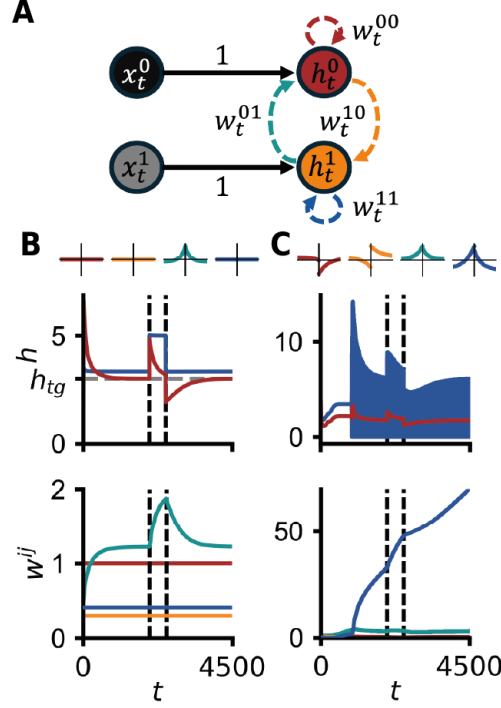
A common method to study spike-timing-dependent plasticity is to perform mean-field analysis [33, 6, 11], which assumes a large network of uncorrelated neurons. Under these assumptions, the weight updates in the spiking network eq. (1) become:

$$\left\langle \frac{dw(t)}{dt} \right\rangle = \eta [r_{pre} \alpha_{XY} + r_{post} \beta_{XY} + r_{pre} r_{post} (\kappa_{XY} \tau_{XY}^{post} + \gamma_{XY} \tau_{XY}^{pre})] \quad (25)$$

with r_{pre} and r_{post} the firing rates of the pre- and post-synaptic neurons. This method allowed us to get a rate "equivalent" of each spike-timing dependent rule defined ineq. (1). We embedded the rate-equivalent rule quadruplets in a 2-neuron linear recurrent network (2RNN) undergoing the familiarity detection task (fig. S1). The activities r_E, r_I of the 2RNN, representing the excitatory and inhibitory spiking populations, followed:

$$\begin{cases} r_E(t+1) = w_{ee}(t)r_E(t) - w_{ie}(t)r_I(t) + x_E(t) \\ r_I(t+1) = w_{ei}(t)r_E(t) - w_{ii}(t)r_I(t) + x_I(t) \end{cases} \quad (26)$$

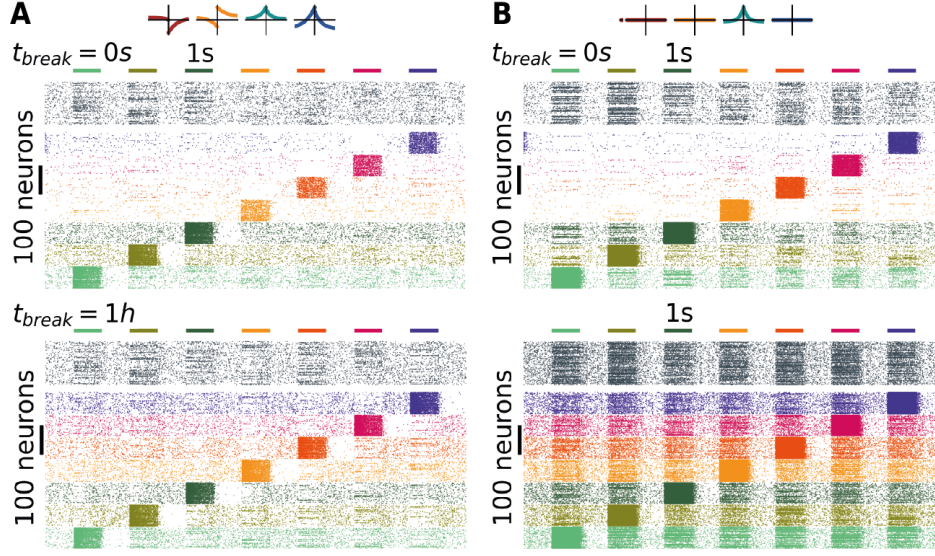
The activities r and weights w were constrained to be positive at all times. The familiarity task was ported to this setting by providing background input to the network $\mathbf{x}_{bg} = (1, 1)$, followed by a training period with stimulus $\mathbf{x}_{stim} = (1.5, 1)$ before reverting to \mathbf{x}_{bg} . Over 95% of the rule quadruplets that were stable in the recurrent spiking model elicited diverging weight or activity dynamics in the rate model, such as the rule quadruplet shown in fig. 1 (fig. S1B). Nevertheless, this 2RNN satisfyingly approximated some rules, particularly those evolving in isolation, such as the rate-equivalent of iSTDP (fig. S1C).



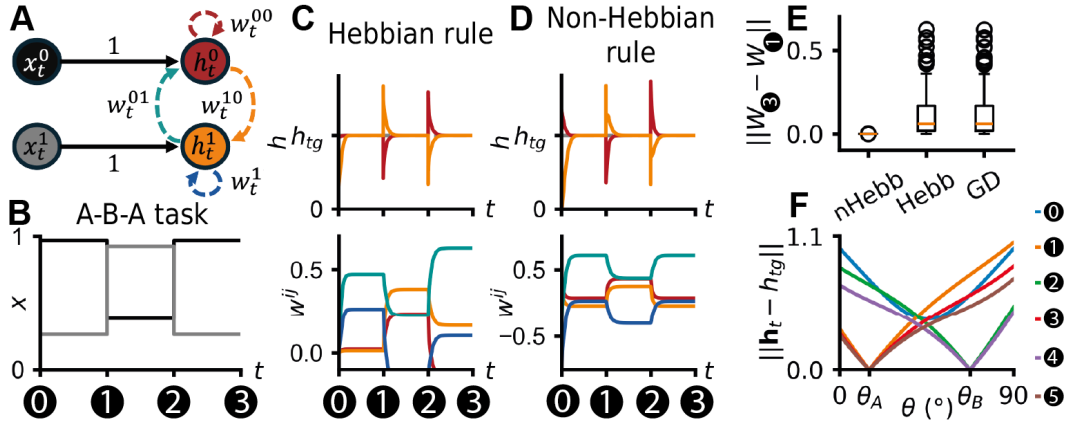
Supplementary Figure S1: **Mean-field-inspired model.** **A:** Linear 2-neuron recurrent network (2RNN) and notations. **B:** 2RNN evolving with the rate-equivalent of the iSTDP rule during the familiarity task (see fig. 2A for spiking equivalent). Dashed lines denote the onset and offset of training. Top: network activities during the task. Bottom: evolution of the four recurrent weights. Colors match the cartoon in A. **C:** Same as B, but for the meta-learned rule quadruplet shown in fig. 1.

Overall this suggested that the assumptions made to obtain the rate equivalent rules were not valid in the case of co-active rules. Indeed, the one rule for which the 2RNN model performed qualitatively similarly like the spiking model is iSTDP, which was shown to decorrelate neuronal activities [11], thus ensuring that the assumption of uncorrelated neuron activities holds. We thus moved to a more abstract setting to understand the memory by accident phenomenon.

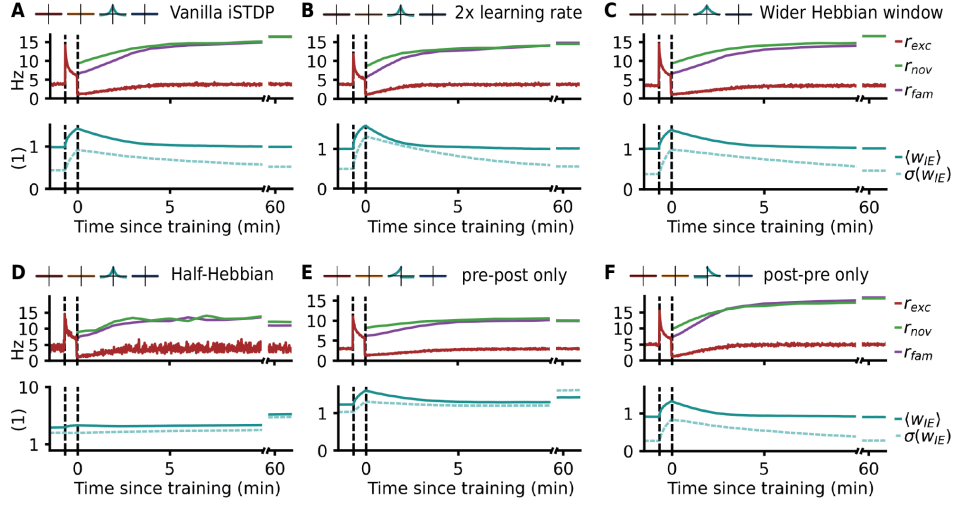
A.6 Supplementary figures



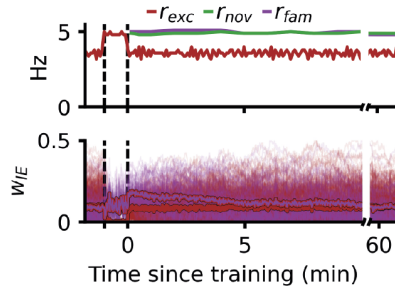
Supplementary Figure S2: **Visualization of network activities for fig. 1.** **A,B:** Raster plots during two test sessions of the familiarity detection task. Neurons are colored by which engram they belong to (see methods for definition of engrams, gray shows neurons not part of any engram).



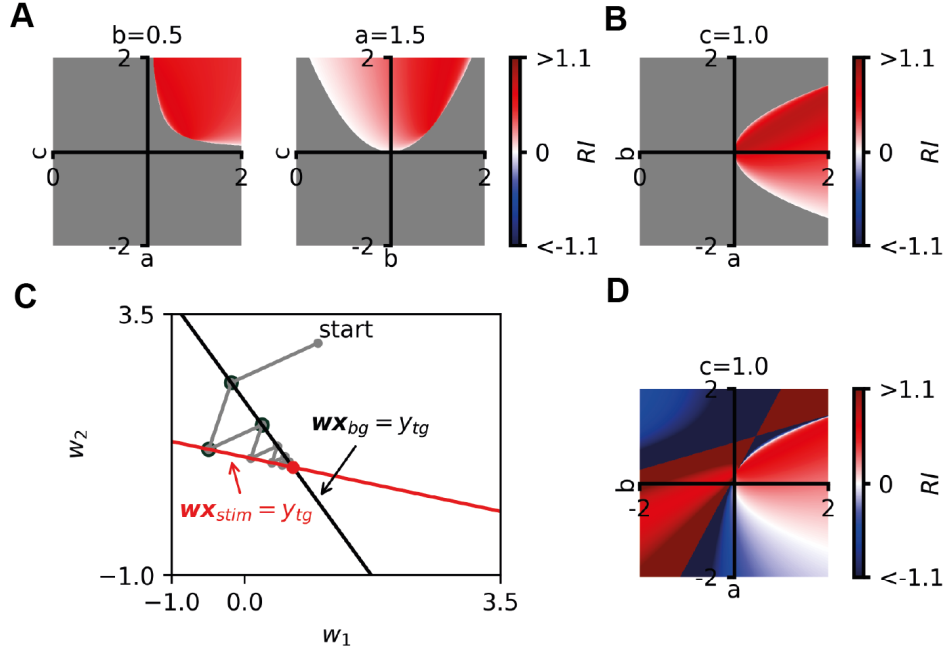
Supplementary Figure S3: **A linear RNN reproduces aspects of memory by accident** **A:** Network and notations, activities are restricted to be positive. **B:** Example input activities in the A-B-A task. Each stimulus presentation is chosen to be long enough for any potential fixed point to be reached. **C:** Hebbian rule in the A-B-A task: $\Delta w_t^{pre post} \propto (h_{tg} - h_t^{post})h_t^{pre}$ with h_{tg} a fixed target (1). Top: Recurrent neurons activities, Bottom: 4 network weights. **D:** Same as C for a non-Hebbian rule: $\Delta w_t^{pre post} \propto (h_{tg} - h_t^{post})$. **E:** Distance between the network weights at the end of the first or the second presentation of stimulus A: averaged over many simulations for the three learning rules tested. "GD" is online gradient descent on the mean squared error loss of the network activity compared to the target activity h_{tg} . **F:** Network response profile for stimuli of various angles at different timepoints of the A-B-A task.



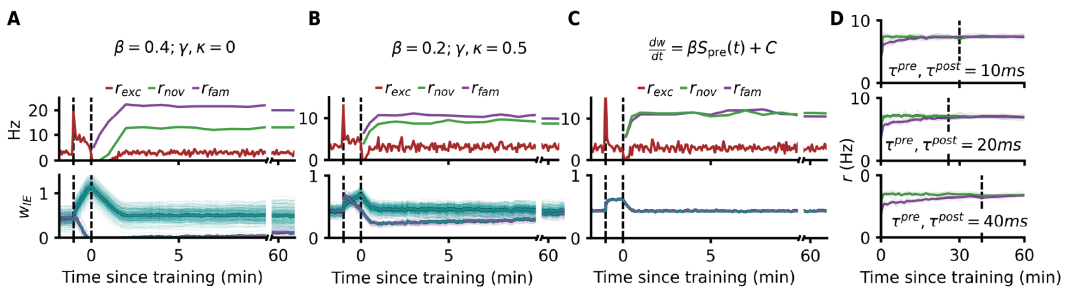
Supplementary Figure S4: **Variations of iSTDP and memory by accident:** Recurrent spiking network undergoing the familiarity detection task, for 6 variants of the iSTDP rule. All variants have the same target excitatory rate of 3Hz. **A:** $\tau_{IE}^{pre} = \tau_{IE}^{post} = 20ms, \alpha_{IE} = -0.12, \beta_{IE} = 0, \kappa_{IE} = \gamma_{IE} = 1$. **B:** $\tau_{IE}^{pre} = \tau_{IE}^{post} = 20ms, \alpha_{IE} = -0.24, \beta_{IE} = 0, \kappa_{IE} = \gamma_{IE} = 2$. **C:** $\tau_{IE}^{pre} = \tau_{IE}^{post} = 40ms, \alpha_{IE} = -0.12, \beta_{IE} = 0, \kappa_{IE} = \gamma_{IE} = 0.5$. **D:** $\tau_{IE}^{pre} = \tau_{IE}^{post} = 20ms, \alpha_{IE} = -0.12, \beta_{IE} = 0.06, \kappa_{IE} = \gamma_{IE} = 0.5$. **E:** $\tau_{IE}^{pre} = \tau_{IE}^{post} = 20ms, \alpha_{IE} = -0.12, \beta_{IE} = 0, \kappa_{IE} = 0, \gamma_{IE} = 1$. **F:** $\tau_{IE}^{pre} = \tau_{IE}^{post} = 20ms, \alpha_{IE} = -0.12, \beta_{IE} = 0, \kappa_{IE} = 1, \gamma_{IE} = 0$.



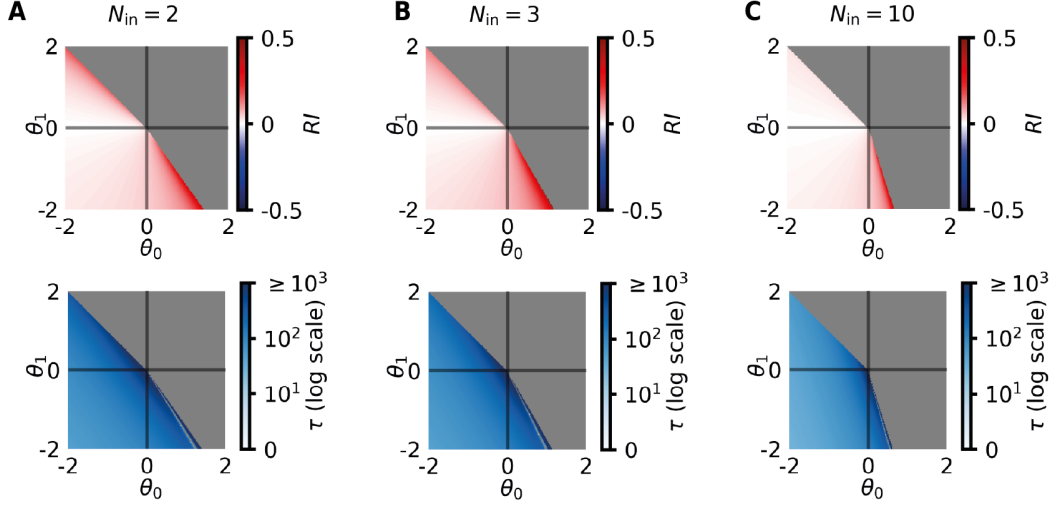
Supplementary Figure S5: **Feedforward spiking mode with E-to-E plasticity.** Top: output neuron firing rate. Bottom: evolution of the 800 excitatory (plastic) weights. Weights from all excitatory input neurons are in red, the subset of weights belonging to the familiar stimulus are overlaid in purple. The means of the two groups (“familiar” weights vs rest) are in bold. The plasticity rule used is $\tau_{EE}^{pre} = \tau_{EE}^{post} = 100ms, \alpha_{EE} = 1, \beta_{EE} = -0.8, \kappa_{EE} = \gamma_{EE} = -1$.



Supplementary Figure S6: **Additional analysis of the implicit feedforward toy model:** **A:** Similar parameter sweeps as in fig. 4B, but varying other parameter combinations. **B:** Same parameter sweep as in fig. 4B, but for an angle of $\frac{\pi}{3}$ between the two inputs. **C:** Grey: dynamics of the $\theta_0 = 0, \theta_1 = -1$ rule from the explicit parameterization in the A-B-A task. Black dots represent the fixed points of the rule associated to $a = 1, b = 0, c = 0$ in the implicit parameterization. **D:** Parameter sweep on the plasticity rules (varying a and b , c fixed. But relaxing the assumption that Σ^{-1} needs to be positive semi-definite).



Supplementary Figure S7: **Additional analysis on testing predictions from toy models.** **A, B, C:** Top: firing rate of the post-synaptic neuron during simulation associated to fig. 5B&C (red), as well as the firing rate in response to the novel and familiar stimuli. Bottom: inhibitory weights, weights from inhibitory input neurons apart of the familiar stimulus are in purple, the rest is in teal. Averages of the two groups are in bold. Dashed lines indicate training onset and offset. **D:** Recurrent spiking network with variants of iSTD (I-to-E plasticity only) in the familiarity task, with different Hebbian windows. Simulations were repeated 5 times, averages across seeds are shown in bold, dashed lines denote the last timestep at which the population firing rate in response to familiar stimuli was significantly different to novel responses (Student t-test, $p < 0.05$).



Supplementary Figure S8: **Extending the explicit toy model to higher dimensions.** **A:** Top: Phase portrait of the relative improvement RI as a function of the values of θ_0 and θ_1 . Note that here, RI is defined as in appendix A.3.2, to extend to $N_{in} > 2$ input dimensions. Grey denotes unstable rules. Bottom: Same parameter sweep as C, but plotting the time to convergence τ . These two plots are generated for $N_{in} = 2$ (same as main). **B,C:** Same as A, but for $N_{in} = 3$ and $N_{in} = 10$.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We show simulations and derivations in four models (recurrent and feedforward spiking networks, two toy feedforward linear models).

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We included a dedicated paragraph in the discussion. We include and discuss negative results from some large scale numerical experiments in the last figure.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings,

model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: For the three parts of this paper that rely on analytical results (mean-field analysis for I-to-E plasticity to compute target firing rates), and the two toy models, we included a dedicated section in appendix for the full derivation.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We included a more detailed method section in Supplementary with the exact parameter values to run our models. The code to reproduce our simulations and analysis is also provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: The dataset for meta-learned rule quadruplets is publicly accessible from previous public work. The code to reproduce all simulations and analysis is available on github.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Supplementary methods go into more details for the exact experiments presented in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: All statistical tests performed are described, and the code provided includes the analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: Paragraph in Supplementary material about the compute-requirements of this work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: this paper is a very fundamental analysis of the emergence of memory in the brain, and has no immediate societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: the models used in this study have no foreseeable potential for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: All libraries upon which the simulations and analysis is built on are cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: a link to the code is provided.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: No human subjects are used in this study.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects were used in this study.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.