

Spontaneous emergence of metacognition in neuronal computation

Hengyuan Ma¹, Wenlian Lu^{1,2,3,4,5,6} and Jianfeng Feng^{1,2,3,4,7,*}

¹*Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China*

²*Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Ministry of Education, Shanghai 200433, China*

³*School of Mathematical Sciences, Fudan University, Shanghai 200433, China*

⁴*Shanghai Center for Mathematical Sciences, Shanghai 200433, China*

⁵*Shanghai Key Laboratory for Contemporary Applied Mathematics, Shanghai 200433, China*

⁶*Key Laboratory of Mathematics for Nonlinear Science, Shanghai 200433, China*

⁷*Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom*



(Received 25 February 2025; accepted 22 July 2025; published 22 August 2025)

Metacognition, a hallmark of human intelligence, enables individuals to assess prediction uncertainty, providing an advantage over artificial intelligence in anticipating risks and performing tasks that demand trustworthiness and reliability. However, the neural mechanisms behind metacognition remain poorly understood. Here, we demonstrate that metacognition can naturally emerge in recurrent neural networks trained on cognitive tasks without guidance from any probabilistic inference rules or additional network architectures. Through naturally embedded nonlinear coupling with the mean of the network output, the covariance of the network output engages in metacognition by assessing the uncertainty associated with the mean, which represents the task responses. We showcase this capability through diverse cognitive tasks and learning algorithms, including reservoir computing and backpropagation. We further propose testable predictions about how key features of neuronal computation in the brain—noise, neuronal correlations, and heterogeneity—contribute to metacognition.

DOI: [10.1103/f1hv-bf1f](https://doi.org/10.1103/f1hv-bf1f)

I. INTRODUCTION

Artificial intelligence (AI) has achieved remarkable progress in various tasks [1]. However, despite their reliance on vast amounts of data and computational power, AI systems still fall short of humanlike intelligence in critical aspects [2]. One key capability humans possess is *metacognition*—the ability to “know about knowing” [3]. Metacognition allows individuals to assess their own cognitive processes by generating a sense of confidence that reflects the likelihood of error in a decision: Greater uncertainty often signals a higher chance of having made a mistake. Moreover, by communicating their uncertainty with cooperators, humans achieve superior performance compared to working alone [4]. While humans can adapt and cooperate by perceiving their cognitive limitations, AI systems often display overconfidence, failing to account for potential errors and risks [5].

Various probabilistic inference techniques have been developed to emulate human metacognition, with the goal of creating more trustworthy and interpretable AI models [5,6]. However, these approaches often require significantly higher computational resources than standard models or rely on complex probabilistic inference rules [7,8], posing a substantial

challenge for scaling to large, real-world systems. In contrast, humans perform metacognitive computations autonomously and with remarkable efficiency. Human metacognition of brain has been widely studied for decades [3,9]. Studies have linked metacognitive abilities to specific brain regions [10–12]. Despite these advances, it remains unclear how neural systems develop metacognitive abilities [13,14]. By uncovering the neural mechanisms underlying human metacognition, we can gain critical insights into how biological systems achieve these capabilities, while also inspiring the development of more reliable AI systems.

Unlike previous studies that either incorporate probabilistic inference rules [15] or require additional network architectures [16] to account for metacognition, the present study demonstrates that metacognitive abilities can spontaneously emerge in a recurrent neural network trained on cognitive tasks, without requiring guidance from any probabilistic inference rules or additional network architectures. We assume that the output of the network is stochastic and can be approximated by a Gaussian distribution $\mathcal{N}(\mu_o, C_o)$, where the mean μ_o encodes the network’s response and the covariance C_o encodes the network’s sense of uncertainty to the response. We trained the network to produce responses that align with the true targets for a cognitive task by supervising the mean μ_o , without explicitly supervising C_o . After training, we found that the uncertainty encoded in C_o reliably predicts the error in μ_o . In other words, the covariance C_o naturally performs metacognition. We then demonstrate how the nonlinear coupling between the mean and covariance of the network gives rise to this metacognitive ability. Furthermore, we provide

*Contact author: jffeng@fudan.edu.cn

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

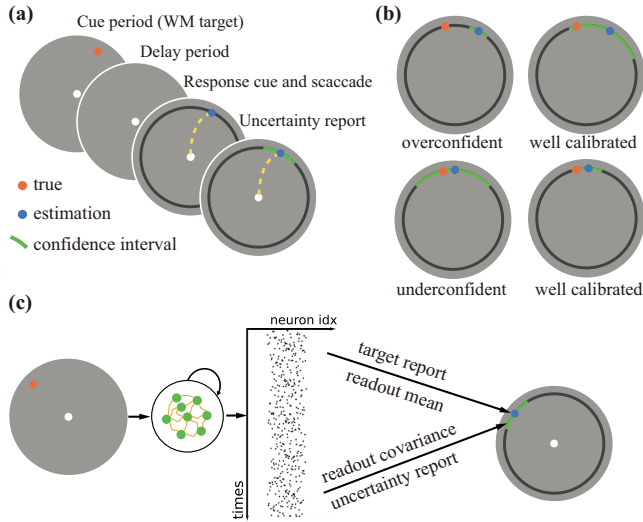


FIG. 1. Metacognition and population coding. (a) Working memory task designed in Ref. [17]. (b) Four typical metacognitive performances. Effective metacognition should accurately reflect the magnitude of the error. (c) A recurrent neuronal network for modeling the (meta)cognition process.

testable predictions about how key features of neuronal computation in the brain—such as noise, neuronal correlations, and heterogeneity—shape (meta)cognition performance, uncovering their beneficial roles.

II. METACOGNITION

We illustrate the concept of metacognition through the working memory (WM) task designed in Ref. [17] [Fig. 1(a)]. Participants are first presented with a cue indicating an angular feature to be memorized. After a delay period, during which the cue is removed, participants report the memorized feature. The cognitive performance is quantified by the deviation between the reported and the true feature, known as the memory error. To assess metacognitive performance, participants also report their uncertainty for each response by adjusting the length of an arc centered at the reported location, which represents a confidence interval for estimating the true feature. As shown in Fig. 1(b), effective metacognition requires uncertainty levels to match error levels. Experimentally, the correlation coefficient between arc length and memory error was relatively high (0.6–0.7; Fig. 5(D) in Ref. [17]), suggesting that participants exhibited strong metacognitive performance.

III. NEURAL MODEL OF METACOGNITION

We modeled the neural processing of a cognitive task using a recurrent neural network of N neurons with a synaptic connection matrix $W \in \mathbb{R}^{N \times N}$. The dynamics of each neuron applied the moment neural network (MNN) [18,19]. MNN represents the neural activity through the mean firing rate $\mu \in \mathbb{R}^N$ and firing covariability $C \in \mathbb{R}^{N \times N}$ [see Eq. (S1) in the Supplemental Material for details [20]]. We assume that (meta)cognitive responses are represented by a population code of the response neurons' activity and can be read out through a linear decoder. Given the network's neural state

(μ, C) , the output is calculated through a linear readout matrix $W_o \in \mathbb{R}^{N_o \times N}$. The output mean $\mu_o = W_o \mu$ encodes the target response, while the output covariance $C_o = W_o C W_o^T$ encodes the level of uncertainty. Throughout this study, we assume that the cognitive task requires the network to generate a continuous periodic feature $\theta \in [0, 2\pi)$, which can be decoded from μ_o . The decoding process and the dimensionality of the output N_o may vary depending on the task.

We trained the network to perform a cognitive task by supervising its output mean μ_o so that $\hat{\theta}$ is close to the target feature θ of training instances and evaluated its (meta)cognitive performance on test instances. Note that only μ_o is supervised, while C_o remains unsupervised. We evaluated the performance through the deviation on the mean, and the uncertainty encoded in the C_o was calculated based on the associated entropy [Eqs. (S5) and (S6) in the Supplemental Material [20]].

IV. METACOGNITION IN WORKING MEMORY

We first investigate metacognition in the WM task, which serves as the foundation for numerous higher-level cognitive processes [21]. In a WM task, the network encodes a variable from a cue and is required to decode the variable after a delay period. WM is prone to errors due to both internal and external noise [22], and significant fluctuations in the error levels have been observed [23–25]. This highlights the demand of a metacognition. Experimentally, humans exhibit excellent metacognitive performance in WM tasks: They are aware of potential errors through a sense of uncertainty [17] and leverage this awareness to optimize their performance [26].

We trained the network to hold WM of a feature θ given the corresponding cue. The response feature is extracted from μ_o through Eq. (S10) in the Supplemental Material [20]. We adapted the reservoir computing approach from Ref. [27], replacing their rate-based neural network with an MNN. Following Ref. [27], we introduced unstructured heterogeneity by adding an untrainable random matrix J with a controllable factor $g > 0$ to the recurrent matrix. The trained network exhibits preferred features through their persistent activities after the cue was removed. Unlike previous modeling of WM where the network's connection matrix is homogeneous and the resulting tuning curves are highly symmetric [28–30], due to the incorporation of the heterogeneity prior to training, the network develops heterogeneous tuning curves and connection weights, as shown in Figs. 2(a)–2(b), aligning with experimental observations [31]. Additionally, each neuron only responds to a narrow range of features, suggesting a sparse coding scheme [32]. Our results complement the findings of Ref. [33]. Whereas their work showed the stability of working memory against small perturbations, we revealed stability of working memory against weight heterogeneity.

Six instances of the network's responses are shown in Fig. 2(c), illustrating significant variability in error levels across instances. This variability aligns with experimental observations of substantial fluctuations in the quality of WM representations within individuals [23,34]. More importantly, larger errors are associated with higher covariance levels, indicating strong metacognitive capabilities. We

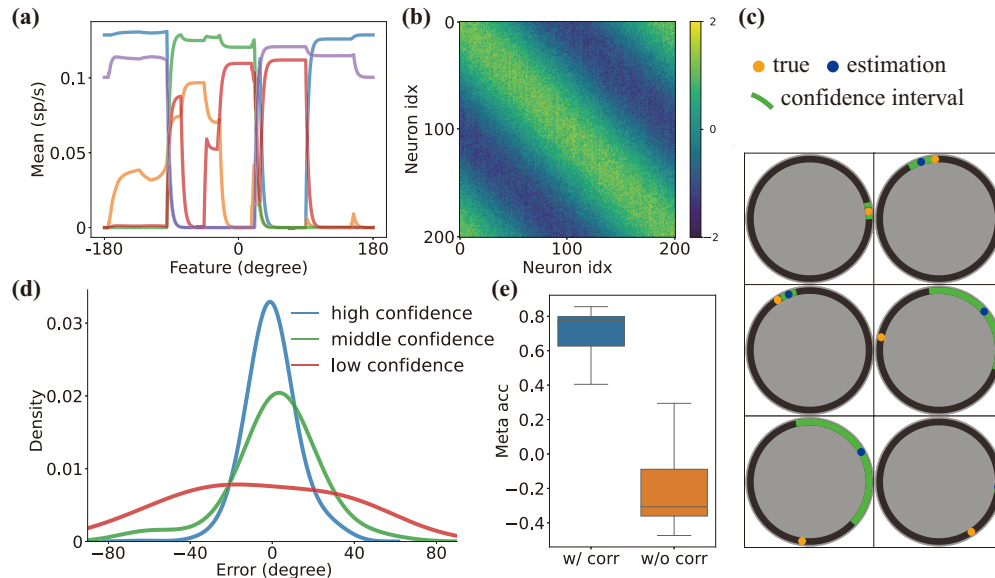


FIG. 2. Metacognition in the working memory task. (a) Tuning curves of five neurons. (b) Trained weights of the network. (c) Six instances of the decoded response and confidence interval. Arc length of the confidence interval is proportional to the square root of the first eigenvalue of the output covariance. (d) Error distribution across three groups of instances, divided based on the uncertainty level. Replication of the experimental results from Fig. 3(b) in Ref. [25]. (e) Metacognition when the correlation between neuron activities is maintained (w/ corr) and when the correlation is clamped to zero (w/o corr). Replication of the experimental results of Fig. 5(D) in Ref. [17].

then divided all test instances into three groups based on uncertainty levels: the low-confidence group (top 25% of uncertainty), the middle-confidence group (25%–50% range of uncertainty), and the high-confidence group (remaining instances). As shown in Fig. 2(d), the low-confidence group exhibits the heaviest-tailed error distribution, followed by the middle-confidence group. This mirrors experimental findings on human introspection during working memory tasks [25,34]. Additionally, the error level aligns with prior studies [17,25,34], and metacognitive accuracy remains high [Fig. 2(e)], consistent with experiments [17].

Next, we examined the role of neuronal correlation in metacognition by clamping all pairwise correlations between neurons to zero and repeated the experiments. We observed a significant decrease in the metacognition accuracy [Fig. 2(e)]. This suggests that neuronal correlation is crucial for metacognition. It aligns with animal experiments that correlated fluctuations between neurons are essential for population coding [35].

V. METACOGNITION IN MULTIPLE COGNITIVE TASKS

We investigated whether metacognition can spontaneously emerge not only in the WM task but also in other cognitive tasks. We examined 16 cognitive tasks that are widely studied in neuroscience, including variants of memory-guided response, context-dependent decision-making, multisensory integration, parametric working memory, and inhibitory control tasks. These tasks take one or two angular features from one or two mods and should respond with an angular feature. Some tasks require the network to memorize stimuli from one modality, while others demand a comparison of stimulus strength across both modalities to generate an output. Additionally, certain tasks involve ignoring stimuli from one

modality based on contextual cues and performing computations accordingly. We adapted the approach that trained a single network through backpropagation to perform all tasks from Ref. [36], replacing their rate-based neural network with an MNN. The response feature is extracted from μ_o through Eq. (S18) in the Supplemental Material [20].

The trained network achieves high accuracy across all tasks, though some tasks show relatively lower accuracy [top panel of Fig. 3(a)]. This variability in performance across tasks with differing levels of difficulty highlights the need for the network to be aware of its errors through metacognition. The metacognition accuracy within each task (green bars) is high, except for tasks where the model achieves perfect accuracy [Fig. 3(a)]. This indicates that the network's covariance spontaneously facilitates metacognition for more challenging tasks. Additionally, the metacognition accuracy across all tasks [red dashed line in the bottom panel of Fig. 3(b)] remains consistently high, indicating that the covariance mechanism can effectively manage errors across diverse cognitive tasks simultaneously. This finding is further corroborated by Fig. 3(b) where error levels and uncertainty levels for each task reveal a strong positive correlation.

More interestingly, we found that the output covariance C_o contains information beyond just uncertainty. As shown in Fig. 3(c), which displays μ_o and C_o for several trial samples of different tasks, we observed that while μ_o 's are quantitatively similar, the corresponding C_o 's vary significantly. This suggests that C_o may carry task-specific information. To test this hypothesis, we trained a convolutional neural network to predict the task type from C_o . The classifier achieved a test set accuracy of 0.647, with a well-structured confusion matrix, as shown in Fig. 3(d), implying that the covariance contains meaningful task-specific information. Notably, downstream neurons may use this information to distinguish

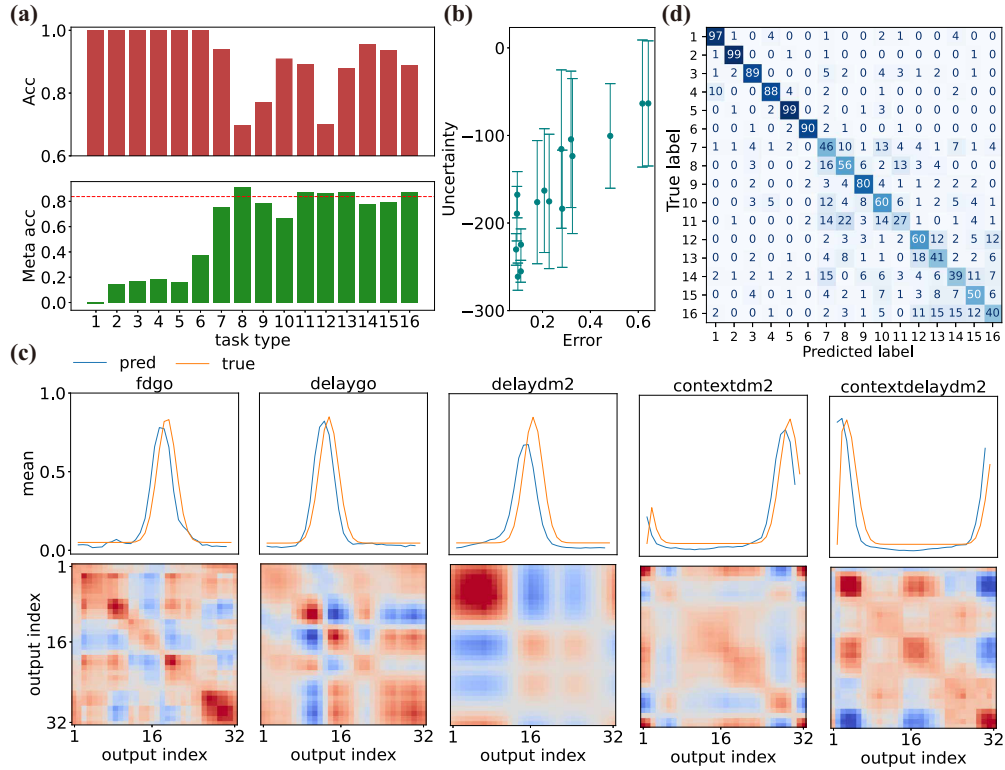


FIG. 3. Metacognition on multiple cognitive tasks. (a) Top: accuracy of each task (see the Supplemental Material for definition [20]). Bottom: metacognition accuracy calculated by the correlation between uncertainty and error within each task. The red dashed line is the metacognition accuracy calculated by the correlation between uncertainty and error across all the tasks. (b) Scatter plot of error and uncertainty level of each task. (c) Output mean μ_o and covariance C_o of several instances of different tasks. (d) Confusion matrix of task-type classification via C_o .

tasks and adapt metacognition. Thus, covariance plays a dual role [37,38], conveying both uncertainty for metacognition and task-specific signals, enabling flexible, context-aware processing.

VI. MECHANISM OF THE METACOGNITION

We analyzed the mechanism of the spontaneous emerging of the metacognition. We first took a deep look into the case of the WM task. As shown in Fig. 4(a), the mean of the network state μ exhibits an imperfect bump, while the covariance C displays a squarelike pattern (each pattern has been centered for comparison, and the three patterns are arranged in increasing order of covariance level). Notably, the bump width (number of neurons with firing rates exceeding 5% of the peak) increases with the covariance level. To further verify this relationship, we calculated the *coupling strength*, defined as the correlation coefficient between bump width and uncertainty metrics [Eq. (S6) in the Supplemental Material [20]]. As shown in Fig. 4(b), the coupling strength is relatively high, indicating a strong coupling between bump width and uncertainty represented by covariance.

Our finding reveals a connection between two leading theories of neural uncertainty representation: probabilistic population coding [39] and sampling-based coding [13]. Probabilistic population coding suggests that both the mean of the target variable and its uncertainty are captured by time-averaged neural responses. In contrast, sampling-based coding

attributes the mean representation to the average response and uncertainty to response variability. While both frameworks successfully link neural activity to behavioral variability observed in experiments, neither fully captures the precise representation of uncertainty required to account for observed errors [17,26,40,41]. Our model admits dual interpretations of uncertainty representation: Since the covariance of neural activity reflects uncertainty, our model can be interpreted as employing a sampling-based coding scheme. Alternatively, as the bump width is strongly positively correlated with the uncertainty metrics [Fig. 4(b)], it can also be used for reflecting the predicted error level of the mean. Therefore, we can interpret the bump width of mean as a direct encoding of uncertainty. Thus, our model implements a form of probabilistic population coding, where both the prediction and its uncertainty are represented by the profile of the mean activity.

We propose that covariance reliably captures uncertainty through a specific mechanism: When errors in the mean are high, the differential covariance component within the covariance increases, amplifying the level of uncertainty. Differential covariance is the component that changes the neural responses across trials along the direction parallel to the tuning curve derivative, hence shrinks its Fisher information [30,42,43]. As illustrated in Fig. 4(c), the differential covariance makes two codes of feature less distinguishable, reducing the coding accuracy. We hypothesize that errors in μ_o arise from random drifts in the bump location [left panel of Fig. 4(d)], which also induce differential covariance. In this

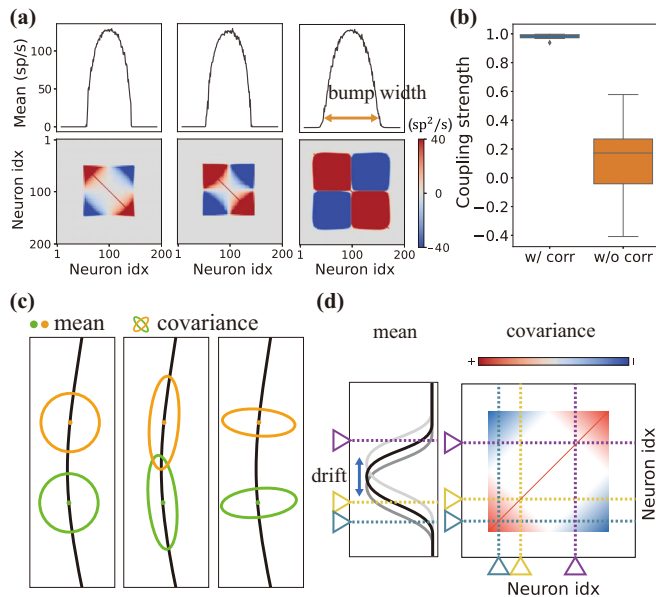


FIG. 4. Mechanism of metacognition for working memory. (a) Typical fixed-point patterns of mean (top) and covariance (bottom), with the bump width increasing from left to right. (b) Coupling strength and bump width when the correlation between neuron activities is maintained (w/ corr) and when the correlation is clamped to zero (w/o corr). (c) Illustration of differential covariance's effect on neural coding: (left) uncorrelated covariance; (middle) tuning curve derivative aligned with principal direction of differential covariance; and (right) tuning curve derivative orthogonal to principal direction of differential covariance. (d) Random drift of bump generates differential covariance components.

structure, neurons on the same side of the bump are positively correlated, while those on opposite sides are negatively correlated [right panel of Fig. 4(d)]. Consequently, higher coding errors are associated with increased differential covariance components. To verify this hypothesis, we introduce an indicator called the differential covariance ratio (DCR) to quantify the amount of differential covariance components (Eq. (S19) in the Supplemental Material [20]). We calculated the correlation between DCR and the uncertainty, which turned out to be $0.951 (\pm 0.02)$. This suggests that differential covariance is the primary source of uncertainty, hence supporting our hypothesis.

We now interpret metacognition in general cognitive tasks beyond working memory. During training, deviations of μ from the true response generate error signals that shift the output mean's bump location. For well-learned instances, these shifts diminish as deviations decrease. However, suboptimal performance leads to persistent deviations, causing continued bump shifts. Frequent bump movement accumulates differential covariance in C , increasing uncertainty for such instances.

Furthermore, we investigated how two other key characteristics of neuronal computations in the brain—noise and heterogeneity—influence both cognitive and metacognitive performance (Fig. S2 in the Supplemental Material [20]). Notably, our results suggest that these characteristics may actually benefit (meta)cognitive processing. We also explored how other factors, such as network size, training set size, and

regularization, impact metacognitive performance (Fig. S3 in the Supplemental Material [20]).

VII. DISCUSSION

This study generates several testable predictions. First, the neuronal correlation is crucial for metacognition [Fig. 2(e)]. This can be tested by decorrelating neural recordings and assess whether metacognition performance is impaired. Second, (meta)cognition may benefit from learning under noisy conditions [Figs. S3(a)–S3(c) in the Supplemental Material [20]]. This could be tested by designing experiments where subjects learn a task under varying levels of noise. Third, network heterogeneity could improve metacognition and coding precision [Figs. S2(g)–S2(i) in the Supplemental Material [20]]. Fourth, the covariance of neural activity may contain task-specific information. This can be tested by assessing whether the covariance of recorded neuronal signals during different cognitive tasks can predict the task types.

Our mechanism analysis suggests that the representation of uncertainty via covariance arises from the coupling between the mean and covariance, whereby errors in the output mean are transmitted to the output covariance through differential covariance. This mechanism implies a potential pathway by which neural populations can develop introspective capabilities. Here, introspective refers to the system's ability to discern what it has learned well versus what it has learned poorly. Furthermore, our findings indicate that differential covariance is not merely a harmful fluctuation that degrades coding accuracy; rather, it may carry informative signals related to task-specific uncertainty and thus serve a functional role. Further investigation is needed to determine the conditions under which differential covariance emerges and how it is decoded by downstream neurons.

We demonstrate that uncertainty representation through output covariance emerges within recurrent MNNs trained on a range of cognitive tasks. While its robustness across tasks suggests a potentially general computational property, whether this capability extends to other architectures—such as feedforward networks—or to different forms of stochasticity and task demands remains an open question for future research.

Several previous works [44–47] implemented neural sampling for representing a given distribution through stochastic neural networks. There are two key distinctions between our work and these studies. First, while prior work focuses on solving probabilistic inference problems using fixed-weight networks, our study involves learning the network weights through training. Second, our model learns to represent the uncertainty of its own predictions—capturing an unknown distribution that depends on prediction quality of the learned network, whereas these studies aim to represent a prespecified distribution. While Ref. [48] showed that network stochasticity reduces prediction error, we specifically examine how it enables uncertainty representation to capture prediction errors. Like Ref. [49], we model the first two moments of neural dynamics. However, their rate-based network does not extend to SNNs (spiking neural networks), whereas our MNN transfers seamlessly to SNNs while preserving key phenomena (Fig. S1 in the Supplemental Material [20]). Moreover,

unlike their approach—which explicitly supervises output covariance during training to fit a ground-truth distribution for uncertainty representation—our model’s metacognition emerges spontaneously without covariance supervision.

In machine learning, uncertainty quantification remains a significant challenge [50,51], with a major obstacle being the implementation of Bayesian inference in the high-dimensional parameter spaces of neural networks. Although methods such as deep ensembles [7] and variational Bayesian inference [8] have been developed, they often incur substantial computational costs. In contrast, humans seem to perform metacognitive evaluations of uncertainty efficiently—though the extent to which these computations align with normative Bayesian principles remains a topic of ongoing debate [52,53]. In neuroscience, studies on uncertainty representation have predominantly focused on data uncertainty, typically modeled through Bayesian frameworks as posterior distributions over unknown parameters, which decrease with accumulating observations. By contrast, representing model uncertainty—arising from limitations in internal models or learning processes—is more complex. Active inference has been proposed as a promising framework to account for such uncertainty, though its role and validity remain actively debated [54,55].

Our study links cognitive phenomena to neuronal computation, identifying two key mechanisms for metacognition: stochasticity (enabling uncertainty assessment via covariance) and population coding (capturing mean deviations through differential covariance). Crucially, we show how metacognition emerges spontaneously, unlike conventional AI approaches that explicitly optimize output covariance [56] yet remain prone to overconfidence [5]. Our findings provide critical insights into how neural systems autonomously achieve effective metacognition through intrinsic computational properties.

ACKNOWLEDGMENT

This work was supported by STI2030-Major Projects 2021ZD0200204, the Science & Technology Commission of Shanghai Municipality (No. 23JC1400800), and the Ji Hua Laboratory S&T Program (No. X250881UG250).

DATA AVAILABILITY

The data that support the findings of this article are openly available [57].

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, GPT-4 technical report, [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [2] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, Sparks of artificial general intelligence: Early experiments with GPT-4, [arXiv:2303.12712](https://arxiv.org/abs/2303.12712).
- [3] S. M. Fleming, Metacognition and confidence: A review and synthesis, *Annu. Rev. Psychol.* **75**, 241 (2024).
- [4] B. Bahrami, K. Olsen, P. E. Latham, A. Roepstorff, G. Rees, and C. D. Frith, Optimally interacting minds, *Science* **329**, 1081 (2010).
- [5] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, A survey of uncertainty in deep neural networks, *Artif. Intell. Rev.* **56**, 1513 (2023).
- [6] D. Seuß, Bridging the gap between explainable AI and uncertainty quantification to enhance trustability, [arXiv:2105.11828](https://arxiv.org/abs/2105.11828).
- [7] S. Fort, H. Hu, and B. Lakshminarayanan, Deep ensembles: A loss landscape perspective, [arXiv:1912.02757](https://arxiv.org/abs/1912.02757).
- [8] A. Graves, Practical variational inference for neural networks, *Adv. Neural Inf. Process. Syst.* **24**, 2348 (2011).
- [9] M. A. Peters, T. Thesen, Y. D. Ko, B. Maniscalco, C. Carlson, M. Davidson, W. Doyle, R. Kuzniecky, O. Devinsky, E. Halgren *et al.*, Perceptual confidence neglects decision-incongruent evidence in the brain, *Nat. Hum. Behav.* **1**, 0139 (2017).
- [10] S. M. Fleming, R. S. Weil, Z. Nagy, R. J. Dolan, and G. Rees, Relating introspective accuracy to individual differences in brain structure, *Science* **329**, 1541 (2010).
- [11] S. M. Fleming, J. Huijgen, and R. J. Dolan, Prefrontal contributions to metacognition in perceptual decision making, *J. Neurosci.* **32**, 6117 (2012).
- [12] A. Kepecs, N. Uchida, H. A. Zariwala, and Z. F. Mainen, Neural correlates, computation and behavioural impact of decision confidence, *Nature (London)* **455**, 227 (2008).
- [13] G. Orbán, P. Berkes, J. Fiser, and M. Lengyel, Neural variability and sampling-based probabilistic representations in the visual cortex, *Neuron* **92**, 530 (2016).
- [14] A. Pouget, J. Drugowitsch, and A. Kepecs, Confidence and certainty: Distinct probabilistic quantities for different goals, *Nat. Neurosci.* **19**, 366 (2016).
- [15] L. S. Geurts, J. R. Cooke, R. S. van Bergen, and J. F. Jehee, Subjective confidence reflects representation of Bayesian probability in cortex, *Nat. Hum. Behav.* **6**, 294 (2022).
- [16] A. Insabato, M. Pannunzi, E. T. Rolls, and G. Deco, Confidence-related decision making, *J. Neurophysiol.* **104**, 539 (2010).
- [17] H.-H. Li, T. C. Sprague, A. H. Yoo, W. J. Ma, and C. E. Curtis, Joint representation of working memory and uncertainty in human cortex, *Neuron* **109**, 3699 (2021).
- [18] J. Feng, Y. Deng, and E. Rossoni, Dynamics of moment neuronal networks, *Phys. Rev. E* **73**, 041906 (2006).
- [19] W. Lu, E. Rossoni, and J. Feng, On a Gaussian neuronal field model, *NeuroImage* **52**, 913 (2010).
- [20] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/f1hv-bf1f> for further simulation details and additional results.
- [21] P. M. Bays, S. Schneegans, W. J. Ma, and T. F. Brady, Representation and computation in visual working memory, *Nat. Hum. Behav.* **8**, 1016 (2024).
- [22] K. Oberauer, S. Lewandowsky, E. Awh, G. D. Brown, A. Conway, N. Cowan, C. Donkin, S. Farrell, G. J. Hitch, M. J. Hurlstone *et al.*, Benchmarks for models of short-term and working memory, *Psychol. Bull.* **144**, 885 (2018).

- [23] R. van den Berg, H. Shin, W.-C. Chou, R. George, and W. J. Ma, Variability in encoding precision accounts for visual short-term memory limitations, *Proc. Natl. Acad. Sci. USA* **109**, 8780 (2012).
- [24] S. Keshvari, R. Van den Berg, and W. J. Ma, Probabilistic computation in human perception under variability in encoding precision, *PLoS ONE* **7**, e40216 (2012).
- [25] R. L. Rademaker, C. H. Tredway, and F. Tong, Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory, *J. Vision* **12**, 21 (2012).
- [26] M. Honig, W. J. Ma, and D. Fougny, Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions, *Proc. Natl. Acad. Sci. USA* **117**, 8391 (2020).
- [27] R. Darshan and A. Rivkind, Learning to represent continuous variables in heterogeneous neural networks, *Cell Rep.* **39**, 110612 (2022).
- [28] Y. Burak and I. R. Fiete, Fundamental limits on persistent activity in networks of noisy neurons, *Proc. Natl. Acad. Sci. USA* **109**, 17645 (2012).
- [29] K. Wimmer, D. Q. Nykamp, C. Constantinidis, and A. Compte, Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory, *Nat. Neurosci.* **17**, 431 (2014).
- [30] H. Ma, Y. Qi, P. Gong, J. Zhang, W.-L. Lu, and J. Feng, Self-organization of nonlinearly coupled neural fluctuations into synergistic population codes, *Neural Comput.* **35**, 1820 (2023).
- [31] C. Langdon, M. Genkin, and T. A. Engel, A unifying perspective on neural manifolds and circuits for cognition, *Nat. Rev. Neurosci.* **24**, 363 (2023).
- [32] M. Beyeler, E. L. Rounds, K. D. Carlson, N. Dutt, and J. L. Krichmar, Neural correlates of sparse coding and dimensionality reduction, *PLoS Comput. Biol.* **15**, e1006908 (2019).
- [33] Á. Ságodi, G. Martín-Sánchez, P. Sokol, and M. Park, Back to the continuous attractor, *Adv. Neural Inf. Process. Syst.* **37**, 66856 (2025).
- [34] D. Fougny, J. W. Suchow, and G. A. Alvarez, Variability in the quality of visual working memory, *Nat. Commun.* **3**, 1229 (2012).
- [35] P. Berkes, G. Orbán, M. Lengyel, and J. Fiser, Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment, *Science* **331**, 83 (2011).
- [36] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, and X.-J. Wang, Task representations in neural networks trained to perform many cognitive tasks, *Nat. Neurosci.* **22**, 297 (2019).
- [37] T. Akam and D. M. Kullmann, Oscillatory multiplexing of population codes for selective communication in the mammalian brain, *Nat. Rev. Neurosci.* **15**, 111 (2014).
- [38] V. C. Caruso, J. T. Mohl, C. Glynn, J. Lee, S. M. Willett, A. Zaman, A. F. Ebihara, R. Estrada, W. A. Freiwald, S. T. Tokdar *et al.*, Single neurons may encode simultaneous stimuli by switching between activity patterns, *Nat. Commun.* **9**, 2715 (2018).
- [39] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget, Bayesian inference with probabilistic population codes, *Nat. Neurosci.* **9**, 1432 (2006).
- [40] D. Fougny, A. Kanabar, T. Brady, and G. Alvarez, Using a betting game to directly reveal the rich nature of visual working memories, *J. Vision* **15**, 1290 (2015).
- [41] A. H. Yoo, L. Acerbi, and W. J. Ma, Uncertainty is maintained and used in working memory, *J. Vision* **21**, 13 (2021).
- [42] A. Kohn, R. Coen-Cagli, I. Kanitscheider, and A. Pouget, Correlations and neuronal population information, *Annu. Rev. Neurosci.* **39**, 237 (2016).
- [43] S. Panzeri, M. Moroni, H. Saffari, and C. D. Harvey, The structures and functions of correlations in neural population codes, *Nat. Rev. Neurosci.* **23**, 551 (2022).
- [44] C. Savin and S. Denève, Spatio-temporal representations of uncertainty in spiking neural networks, *Adv. Neural Inf. Process. Syst.* **27**, 2024 (2014).
- [45] P. Masset, J. Zavatone-Veth, J. P. Connor, V. Murthy, and C. Pehlevan, Natural gradient enables fast sampling in spiking neural networks, *Adv. Neural Inf. Process. Syst.* **35**, 22018 (2022).
- [46] Y. Qi and P. Gong, Fractional neural sampling as a theory of spatiotemporal probabilistic computations in neural circuits, *Nat. Commun.* **13**, 4572 (2022).
- [47] W.-H. Zhang, S. Wu, K. Josić, and B. Doiron, Sampling-based Bayesian inference in recurrent circuits of stochastic spiking neurons, *Nat. Commun.* **14**, 7074 (2023).
- [48] C. E. Rullán Buxó and J. W. Pillow, Poisson balanced spiking networks, *PLoS Comput. Biol.* **16**, e1008261 (2020).
- [49] R. Echeveste, L. Aitchison, G. Hennequin, and M. Lengyel, Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference, *Nat. Neurosci.* **23**, 1138 (2020).
- [50] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Inf. Fusion* **76**, 243 (2021).
- [51] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, *Adv. Neural Inf. Process. Syst.* **32**, 13991 (2019).
- [52] H.-H. Li and W. J. Ma, Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis, *Nat. Commun.* **11**, 2004 (2020).
- [53] J. Navajas, C. Hindocha, H. Foda, M. Keramati, P. E. Latham, and B. Bahrami, The idiosyncratic nature of confidence, *Nat. Hum. Behav.* **1**, 810 (2017).
- [54] K. Friston, The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **11**, 127 (2010).
- [55] M. Colombo and C. Wright, First principles in the life sciences: The free-energy principle, organicism, and mechanism, *Synthese* **198**, 3463 (2021).
- [56] A. Kendall and Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **30**, 5574 (2017).
- [57] H. Ma, W. Lu, and J. Feng, mnn-metacognition-codes, 2025, https://github.com/AwakerMhy/mnn_wm_uq.