Supporting Text
Neural Complexity

The neural complexity C_N of a system is defined as the ensemble average mutual information (MI) between subsets of a given size, ranging from 1 to $n/2$ for a neural system composed of n elements, and their complements (1). This value is high if each of the system's subsets can take on many different states and if these states make a difference to the rest of the system. In practice, C_N of a network X is calculated as follows:

$$C_N(X) = \sum_{k=1}^{n_t/2} \langle \mathrm{MI}(X_j^k; X - X_j^k) \rangle,$$

where

$$\mathrm{MI}(A; B) = H(A) + H(B) - H(AB).$$

In the case of Gaussian-distributed system variables, the entropy calculation simplifies to:

$$H(X) = \frac{1}{2} \ln((2\pi e)^n |\mathrm{COV}(X)|).$$

In the above, k is the subset size, X_j^k is the j-th subset of size k, n_t is the total number of subsets of size k, $\langle\rangle$ is the ensemble average, $\mathrm{MI}(A; B)$ is the mutual information between A and B, $H(X)$ is the entropy of X, $\mathrm{COV}(X)$ is the covariance matrix of X, and $|.|$ indicates the matrix determinant. An approximation to C_N which considers only bipartitions consisting of a single element and the rest of the system (2) is given by:

$$C(X) = H(X) - \sum_{k=1}^{n} H(x_i | X - x_i),$$

where $H(A|B)$ is the conditional entropy of A given B. For reviews of neural complexity and associated measures, see refs. 3-5 .

Information Integration

Information integration, or Φ, is proposed to measure the total amount of information that a system can integrate. It is calculated as follows (see refs. 6 and 7 for a full account): Given a system of n elements, one identifies all possible bipartitions of the system. For each bipartition $A|B$, one replaces the outputs from A by uncorrelated noise (i.e., maximally entropic activity), and one measures how differentiated are the responses of its complement (B). This is the effective information (EI) between A and B:

$$\mathrm{EI}(A{\rightarrow}B) = \mathrm{MI}(A_{Hmax}; B),$$

where $\mathrm{MI}(A_{Hmax}; B)$ is the mutual information between A and B when the outputs from A have maximal entropy. $\mathrm{EI}(A{\rightarrow}B)$ measures the capacity for causal influence of partition A on its complement B (i.e., all possible effects of A on B). Given that $\mathrm{EI}(A{\rightarrow}B)$ and $\mathrm{EI}(B{\rightarrow}A)$ are not necessarily equal, one can define:

$$\mathrm{EI}(A{\leftrightarrow}B) = \mathrm{EI}(A{\rightarrow}B) + \mathrm{EI}(B{\rightarrow}A).$$

The minimum information bipartition (MIB) is the bipartition for which the normalized $\mathrm{EI}(A{\leftrightarrow}B)$ is lowest. Normalization is accomplished by dividing $\mathrm{EI}(A{\leftrightarrow}B)$ by $\min\{H_{max}(A); H_{max}(B)\}$, so that effective information is bounded by the maximum entropy available. The resulting MIB corresponds to the informational "weakest link" of the system, and the Φ value of the system is the nonnormalized $\mathrm{EI}(A{\leftrightarrow}B)$ across the MIB.

A further stage of analysis has been described (7) in which a system can be decomposed into "complexes" by calculating Φ for different subsets of elements; a complex is a subset having $\Phi > 0$ that is not included in a larger subset with higher Φ. As we show below, the example Hopfield-like network described in the main text cannot be decomposed into complexes because the Φ value of the complete network is greater than that of any subset.

Determining Φ for a Hopfield-Type Network. Let us consider a simple fully connected Hopfield network (8),

$$x_i(t+1) = \sum_{j=1}^{n} 2^j f(x_j(t)) \,, \qquad \text{where} \quad f(x) = \begin{cases} -1 & \text{if } x < 0 \\ +1 & \text{if } x \geq 0 \,, \end{cases} \qquad [\mathbf{1}]$$

where each $x_i(t)$ denotes the state of neuron i at time t, and $n \geq 1$ is the number of neurons. The synaptic connections among the neurons are chosen

in such a way that the effect of neuron $x_j(t)$ on every other neuron in the network (including itself) is $2^j f(x_j(t))$, i.e., $\pm 2^j$. Apparently, the dynamics of this network is driven by the last neuron $x_n(t)$.

Theorem 1. *The Hopfield network* [Eq. **1**] *of size n has information integration*

$$\Phi = n \text{ bits.}$$

The proof of the theorem is based on a number of propositions that are stated and proved below. For clarity, we treat Eq. **1** as a discrete-time dynamical system, and we refer to "neurons" as elements.

First, let us establish boundary conditions on the activities of elements in Eq. **1**:

Proposition 1. *Let*

$$a = \sum_{j=1}^{n} 2^j = 2^{n+1} - 2 \ ,$$

then all $|x_i(t)| \leq a$ *for all* $t > 0$.

Proof. The maximal (minimal) value of the sum in Eq. **1** is obtained when all $x_j(t)$ are positive (negative), so that $f(x_j(t)) = +1$ (-1). □

Let us now show that the network [Eq. **1**] has very simple behavior in which all elements receive identical inputs, and hence have identical states:

Proposition 2. *The dynamical system* [Eq. **1**] *has only two attractors*

$$x_1 = x_2 = \cdots = x_n = \pm a \ ,$$

where a is defined above. It converges to one or the other attractor depending on the sign of $x_n(0)$. *The convergence takes only two time steps.*

Proof. Because $2^n > \sum_{j=1}^{n-1} 2^j = 2^n - 2$ (Proposition 1), the sign of every variable, $x_i(t+1)$, in Eq. **1** is determined completely by the term $2^n f(x_n(t))$.

Suppose $x_n(0) \geq 0$ (so that $f(x_n(0)) = +1$), then the first iteration results in all $x_i(1) > 0$. Because all $f(x_j(1)) = +1$, the second and all subsequent iterations result in $x_i(t) = \sum_{j=1}^{n} 2^j = a$, for all $t = 2, 3, \ldots$. Similarly, if $x_n(0) < 0$, then all $x_i(1) < 0$, and $x_i(t) = -a$ for all $t = 2, 3, \ldots$. □

Let us show now that the transfer function in Eq. **1** maps different state vectors of the network into different integer values:

Proposition 3. *Let A be a subset of $\{1, 2, \ldots, n\}$ with $k \leq n$ distinct elements. Let y_j, $j \in A$, be k variables from the domain $\{-1, +1\}$. The function F defined by:*

$$F(y) = \sum_{j \in A} 2^j y_j \ , \qquad y = \{y_j\}, j \in A \ , \qquad \textbf{[2]}$$

maps different vectors y to different integer numbers, i.e., it is injective (an embedding).

Proof. Let us use a new coordinate system $z_j = (1 + y_j)/2$ if $j \in A$ and $z_j = 0$ if $j \notin A$. Apparently, $z_j = 0$ or $+1$. In the new coordinates, the function F has the form

$$F(z) = 2 \left(\sum_{j=1}^{n} 2^j z_j \right) - \sum_{j \in A} 2^j$$

The term in parenthesis is an integer number with binary representation $z = (z_1, \ldots, z_n)$, which is determined uniquely by the vector y. The other terms are constants. Hence, different vectors y (corresponding to binary vectors z) result in different integer numbers defined by Eq. **2**. □

Let us now derive the entropy of the output from a subset A in Eq. **1** in the case where the output from each element in A is replaced by random noise:

Proposition 4. *Let A be a subset of $\{1, 2, \ldots, n\}$ with $k \leq n$ distinct elements. Let $y_j(t)$, $j \in A$, be k independent random variables that take values $\{-1, +1\}$ with probabilities 0.5. The entropy of the random variable*

$$\sum_{j \in A} 2^j y_j(t)$$

is k bits.

Proof. The vector $y(t)$ can assume 2^k different values with equal probability. Indeed, it is of the form $(\pm 1, \pm 1, \ldots, \pm 1)$ with k elements. From Proposition 3 it follows that the random variable also assumes 2^k distinct random values, which have equal probabilities $p = 2^{-k}$. Hence, the entropy is:

$$H = -\sum_{j \in A} p \log p = -2^k (2^{-k} \log 2^{-k}) = k \ ,$$

which follows from the Shannon−Weaver formula (9) . We measure entropy in bits so that logarithms are taken with base 2 and $\log 2 = 1$. □

Let us now derive the entropy of the output from the entire network [Eq. **1**] in the case where the output of each element in A is replaced by random noise:

Proposition 5. *Let A be a subset of $\{1, 2, \ldots, n\}$ with $k < n$ elements. Let y_j, $j \in A$, be k random variables $y_j(t)$ that take values $\{-1, +1\}$ with probabilities* 0.5*. Each random variable defined by (see* Eq. **1***):*

$$x_i(t+1) = \sum_{j \in A} 2^j y_j(t) + \sum_{j \notin A} 2^j f(x_j(t)), \quad i \in (1, \ldots, n), \qquad \textbf{[3]}$$

has the following entropy:

- $H = k$, *if $n \notin A$, or*
- $H = k + 1$, *if $n \in A$.*

Proof. Suppose $n \notin A$, then all $x_i(t+1)$, $i \notin A$, have the same sign as $x_n(t)$, which has the same sign as $x_n(0)$. Then, Eq. **3** can be rewritten in the form:

$$x_i(t+1) = \sum_{j \in A} 2^j y_j(t) + \text{sign}\,(x_n(0)) \sum_{j \notin A} 2^j \ . \qquad \textbf{[4]}$$

Note that the second term of Eq. **4** is a constant. From Proposition 4 it follows that the entropy of the random variable $x_i(t)$ is k.

Now suppose that $n \in A$. Then the sign of all $x_j(t)$, $j \notin A$, is equal to the sign of $y_n(t-1)$. That is, Eq. **3** has the form:

$$x_i(t+1) = \sum_{j \in A} 2^j y_j(t) + y_n(t-1) \sum_{j \notin A} 2^j \ . \qquad [5]$$

Because $y_n(t-1)$ is independent from $y_n(t)$, the entropy of this sum is $k+1$. □

Now we need to derive the effective information between two nonoverlapping subsets, A and B, of Eq. **1** in the case when one subset, for example A, is replaced by random noise. For this case, we need to describe the state of the union AB, and there are at least two alternative ways to do that, as shown in Fig. 1. We note that the definition of Φ (in ref. 7) does not consistently specify which alternative to use. For the purpose of illustration, we assume that each set consists of only one element, e.g., $A = \{1\}$ and $B = \{2\}$. Then, the state of the union $AB = \{1, 2\}$ can be treated as

- ALTERNATIVE a: $(\tilde{x}_1(t), x_2(t))$, i.e., the variable $x_1(t)$ is replaced by the random variable $\tilde{x}_1(t)$, so that the output $y_1(t) = f(\tilde{x}_1(t))$ will be transmitted when the time variable steps from t to $t+1$ (in other words, the random output at time t will exert influence on $\{2\}$ at time $t+1$).
- ALTERNATIVE b: $(x_1(t), x_2(t))$, i.e., the replacement of the output from $x_1(t)$ by the random variable $y_1(t)$ occurs somewhere along the transmission line.

The first alternative implies that the state of A at time t is completely independent from the state of B, so that the entropy $H(AB) = H(A) + H(B)$, and the effective information $\mathrm{EI}(A \to B) = \mathrm{MI}(A_{Hmax}; B) = H(A) + H(B) - H(AB) = 0$, not only for the Hopfield network above, but for any discrete-time dynamical system. This alternative therefore leads to the undesirable conclusion that $\Phi = 0$ for all such systems. Thus, we use the second alternative in the theorem below.

Proposition 6. *Let A and B be two distinct nonoverlapping subsets of the set $\{1, 2, \ldots, n\}$. Let k be the number of elements in A. Let us replace the*

outputs from the elements in the subset A in Eq. **1** *by random variables that maximize the entropy $H(A)$. Then, the effective information from elements in A to elements in B in Eq.* **1**, *defined as:*

$$\mathrm{EI}(A \to B) = \mathrm{MI}(A_{Hmax}; B) = H(A) + H(B) - H(AB) \ ,$$

is k.

Proof. The maximal entropy of outputs from A is $H(A) = k$ when each output is replaced by a random variable $y_j(t) = \pm 1$ with the probability 0.5 of being positive or negative, so we can use Proposition 5.

Suppose first that $n \notin A$, then from Proposition 5 it follows that every element in B has the entropy k. Because elements of B receive identical inputs, they have identical states, so the entropy of the whole of B is also k. Moreover, from Eq. **4** it follows that the state of B is a linear combination of only y_j, $j \in A$. All elements in the union AB also receive identical inputs, hence they have identical states, and the entropy of the union is the entropy of any constituent element, hence $H(AB) = k$.

Now suppose that $n \in A$. Repeating the same arguments as above, and using Eq. **5**, we conclude that the entropy of each element in B, the entropy of the whole of B, and the entropy of the union AB is $k + 1$.

In both cases (i.e., whether $n \in A$ or not) $\mathrm{EI}(A \to B) = k$. □

Corollary 1. *Let A and B be two distinct nonoverlapping subsets of the set $\{1, 2, \ldots, n\}$. Let m be the total number of elements in the union AB. Then, the effective information between elements in A and B in* Eq. **1**, *defined as:*

$$\mathrm{EI}(A \leftrightarrow B) = \mathrm{EI}(A \to B) + \mathrm{EI}(B \to A) \ ,$$

equals m.

Proof. Let k_A be the number of elements in A, and k_B be the number of elements in B, so that $k_A + k_B = m$. From Proposition 6 it follows that $\mathrm{EI}(A \to B) = k_A$ and $\mathrm{EI}(B \to A) = k_B$. □

Proof. Now let us prove Theorem 1. Consider a subset, S, of m elements in the Hopfield network [Eq. **1**]. Let $A|B$ be any bipartition of S. From Corollary 1, it follows that the effective information across this bipartition is

m. Because all bipartitions are equivalent with respect to effective information, the effective information across the minimal information bipartition of S, and therefore the information integration value for the subset S, is also m (we do not even need to find the minimal information bipartition). Because larger subsets have higher values of information integration, information integration is maximal when S is the whole network. Therefore, the information integration value for the network, denoted by Φ, is n, i.e., the network size. $\square$

Φ Depends on the Choice of Variables. Interestingly, the network [Eq. **1**] can be rewritten in a different form:

$$v_i(t+1) = f\left(\sum_{j=1}^{n} 2^j v_j(t)\right), \qquad \textbf{[6]}$$

where each $v_i(t) = f(x_i(t))$ is a new, binary or spike-like, variable that describes the output from the i-th neuron. The two systems are equivalent in the sense that if they start with the same initial conditions, they will produce identical behavior. However, as we show below, the system described by Eq. **6** has $\Phi \leq 2$, a value which does not grow as a function of network size n.

Consider a $(k, n-k)$ bipartition, $A|B$, of the network with the subset A having k elements. Replace the outputs from each neuron v_i in A by either $+1$ or -1 (with probability 0.5), to maximize the entropy of A, which is

$$H(A) = k\ .$$

Because $v = f(\text{input})$, each element in B has the state $v = +1$ or $v = -1$, therefore its entropy is less than or equal to $\log 2$, which is 1 (recall that in x coordinates, this value was equal to k). Because every element in B receives identical input, the entropy of the whole of B,

$$H(B) \leq 1\ .$$

Furthermore, because elements in B are driven by output from A, we can repeat the steps in Proposition 5 to show that the entropy of the whole network is equal to the entropy of the subset A plus the entropy of B at the previous time-step, i.e.,

$$H(AB) \geq k\ .$$

Therefore, the effective information,

$$\mathrm{EI}(A \to B) = \mathrm{MI}(A_{Hmax}; B) = H(A) + H(B) - H(AB) \leq k + 1 - k = 1 \; .$$

Similarly, $\mathrm{EI}(B \to A) \leq 1$, which implies that any bipartition has:

$$\mathrm{EI}(A \leftrightarrow B) = \mathrm{EI}(A \to B) + \mathrm{EI}(B \to A) \leq 2 \; .$$

Because this value does not grow as a function of network size n, neither does the corresponding value for Φ.

The above result reinforces our comment in the main text that all complexity measures or other quantities based on Shannon entropy necessarily depend on the exogenous choice of variables chosen to characterize the system. In the above example, the Φ value can either be bounded by a constant or it can be an unbounded function of system size, depending on the variables selected. According to the information integration theory of consciousness, the appropriate spatial and temporal scales for measurement should be chosen according to the criterion that the corresponding value of Φ is maximized (7). However, because we have shown that Φ can be unbounded even for a simple class of Hopfield-type network, this criterion cannot be applied in practice.

Calculating Φ for Continuous Variables. Consider a system composed of two coupled oscillators A and B with continuous dynamics and having phases ϑ and θ defined on the interval $[0, T]$, where $T > 0$ is the period of oscillation. The coupling is assumed to be strong so that the oscillators quickly synchronize and maintain the in-phase relationship, i.e., $\vartheta = \theta$. The effective information from A to B is the same as from B to A, and it is equal to the entropy of a continuous random variable on the interval $[0, T]$, denoted here as $H([0, T])$. Because the system consists of only two elements, the minimal information bipartition (MIB) is $A|B$, and therefore Φ is equal to $2H([0, T])$. However, the entropy, $H([0, T])$, of a continuous variable is infinite; hence, the system of two oscillators has an infinite value for Φ.

To see why the entropy of a continuous variable is infinite, we can apply the Shannon−Weaver formula (9) to the interval $[0, T]$. Let us subdivide this interval into m subintervals of equal length T/m. Because A and B are oscillators, their phase variables monotonically increase and sweep the

interval $[0, T]$ periodically. Hence, the probability of the phase falling in any subinterval is $p = 1/m$, and the entropy is

$$-\sum p \log p = -\sum \frac{1}{m} \log \frac{1}{m} = \log m.$$

Now, if the subdivisions are refined by making m larger, the entropy increases, and it becomes infinite in the limit $m \to \infty$. One could try to salvage this situation by considering the differential entropy (10) :

$$H([0\ T]) = -\int p(x) \log(p(x)) dx = \log T, \qquad \text{(because } p(x) = 1/T\text{)}.$$

However, differential entropy, by its definition, is dependent on the units of measurement of the relevant variables: here, the units of time used to measure T. Because this dependence is logarithmic, differential entropy can become negative. In the present example, differential entropy would be zero for $T = 1$ and negative for $T < 1$. This creates a problem for the determination of the MIB for a system, because in order to do so, it is necessary to normalize $\mathrm{EI}(A \leftrightarrow B)$ by $\min\{H_{max}(A); H_{max}(B)\}$ (see ref. 7 and above). This normalization is not well defined when the latter quantity is zero or negative. In summary, a simple system characterized by continuous variables has either infinite Φ or requires an exogenous choice of measurement units in order to allow calculation of a well-defined value for Φ.

As we mention in the text, it is expected that quantitative measures of relevant complexity will vary according to the variables chosen to characterize the system, and also according to the units of measurements chosen for these variables. These dependencies apply not only to Φ, but also to neural complexity C_N and to causal density *cd*. However, in contrast to Φ, neither of these alternative measures has been proposed to reflect a fundamental quantity that corresponds to the "amount" of subjective experience.

Causal Density

In previous work, we have developed the use of Granger causality (11) for the analysis of simulated neural systems (12, 13). As we mention in the main text, the causal density (*cd*) of a network's dynamics reflects the fraction of interactions among nodes that are causally significant (12). *cd* is calculated as $\alpha/(n(n-1))$, where α is the total number of significant causal links observed, and n is the number of elements in the network. Because *cd* is dependent only on the proportion of interactions that are causally significant, it is bounded in the range [0,1].

The contribution, or "magnitude" of a causally significant connection can be estimated as the logarithm of the corresponding F statistic (see ref. 14 and main text). This allows the calculation of a "weighted" version of causal density which takes into account the varying contributions of each interaction. Weighted causal density, cd_w, is calculated as:

$$cd_w = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (f(ij)(\tau_{ij}) + f(ji)\tau_{ji})),$$

where τ_{ij} is the magnitude of the causal interaction from element i to element j, and $f(.)$ is equal to 1 if the interaction is statistically significant and 0 otherwise. The two measures *cd* and cd_w are complementary: *cd* emphasizes the distribution of causal interactions throughout a network but is not sensitive to their varying contributions; cd_w takes into account these variations, but its value for a network may in some cases be dominated by a small number of highly significant interactions.

MATLAB (Natick, MA) routines for calculating causal density are provided on the website www.nsi.edu/users/seth/.

References

1. Tononi, G., Sporns, O. & Edelman, G.M. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 5033–5037.

2. Tononi, G., Edelman, G.M. & Sporns, O. (1998) *Trends Cognit. Sci.* **2**, 474–484.

3. Sporns, O. & Tononi, G. (2002) *Complexity* **7**, 28–38.

4. Sporns, O., Tononi, G. & Edelman, G.M. (2000) *Cereb. Cortex* **10**, 127–141.

5. Seth, A.K. & Edelman, G.M. (2004) in *Complex Networks, Lecture Notes in Physics*, eds. Naim, E. Ben, Fraunfelder, H. & Toroczkai, Z. (Springer, Berlin), pp. 487–518.

6. Tononi, G. & Sporns, O. (2003) *BMC Neurosci.* **4**, 31.

7. Tononi, G. (2004) *BMC Neurosci.* **5**, 42.

8. Hopfield, J.J. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 2554–2558.

9. Shannon, C.E. & Weaver, W. (1949) *The Mathematical Theory of Communication* (Univ. of Illinois Press, Urbana, IL).

10. Cover, T.M. & Thomas, J.A. (1991) *Elements of Information Theory.* (Wiley Interscience, New York).

11. Granger, C.W.J. (1969) *Econometrica* **37**, 424–438.

12. Seth, A.K. (2005) *Network: Comput. Neural Sys.* **16**, 35–54.

13. Krichmar, J.L., Seth, A.K., Nitz, D.A., Fleischer, J.G. & Edelman, G.M. (2005) *Neuroinformatics* **3**, 197–222.

14. Geweke, J. (1982) *J. Am. Stat. Assoc.* **77**, 304–13.