

Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework

David Balduzzi - balduzzi@wisc.edu; Giulio Tononi - gtononi@wisc.edu

Supplementary Information

1. Relative entropy and entropy

Relative entropy. Given two probability distributions p and q defined on a set Z of N possible outcomes, the relative entropy [1] of p with respect to q is given by

$$H[p \parallel q] := \sum_{i=1}^N p_i \log_2 \left(\frac{p_i}{q_i} \right).$$

Relative entropy is also known as Kullback-Leibler divergence. It can be shown that relative entropy is always non-negative, $0 \leq H[p \parallel q]$, and further that $H[p \parallel q] = 0$ if and only if $p = q$. A special case arises in the context of Bayesian updating. Suppose $p(H)$ is the prior distribution for a set of hypotheses, and $p(H|D = d)$ is the posterior distribution given some new data d . The relative entropy $H[p(H|D = d) \parallel p(H)]$ quantifies the number of bits of information about H that are gained by learning $D = d$.

We extend the interpretation to arbitrary distributions p and q , where q is a prior distribution and p is a posterior. We take a posterior distribution to be a distribution that is specified after causal interactions have taken place. Relative entropy $H[p \parallel q]$ measures the information generated by specifying p , relative to q . We emphasize that it is meaningless to measure the information generated by a repertoire p without a yardstick, given here by q .

Entropy. It is useful to compare relative entropy to the more familiar notion of (information) entropy. Entropy is defined as

$$H(p) := \sum_i \log_2(p_i),$$

where we use round brackets to distinguish the two entropies. A related notion is the concept of self-information. Given a set of outcomes Z distributed according to p , the self-information given by outcome z_i is $I(z_i) := -\log_2(p_i)$. Self-information is usually interpreted as a measure of the surprise associated with outcome z_i . If $p(z_i) = 1$ then $I(z_i) = 0$, there is no surprise; if $p(z_i) = 0$ then $I(z_i) = \infty$, an impossible outcome is infinitely surprising. Comparing the formulae for entropy and self-information it is clear that entropy is the expected value of the self-information of Z .

We apply this as follows. Let δ_i be the probability distribution on Z assigning probability 1 to outcome z_i and probability 0 to all other outcomes. Observe that $I(z_i) = H[\delta_i \parallel p]$, which we interpret as measuring how much information is generated by the particular outcome z_i from the repertoire of possible outcomes (possible according to distribution p). It follows that

$$H(p) = E_p \left\{ H[\delta_i \parallel p] \right\} = \sum_i p_i \cdot H[\delta_i \parallel p],$$

so entropy is the *expected* information generated by an outcome of repertoire p . Thus entropy can be derived from the more fundamental notion of relative entropy.

2. The maximum entropy distribution on a finite set

The maximum entropy (maxent) distribution on a finite set Z of possible outcomes is the distribution m that maximizes the entropy functional: $m(Z) := \operatorname{argmax}_p H(p)$. If there are N outcomes in Z then the maxent distribution is given by

$$m(z_i) = \frac{1}{N} \text{ for all } z_i \in Z.$$

The maxent distribution on Z coincides with the uniform distribution on Z . In the paper, the set Z will arise as the set of possible outputs of a collection of n elements in a system X . If the elements are all binary (with outputs 0 and 1), then Z consists of $N = 2^N$ possible activity patterns. A useful property of the maxent distribution is that it is independent on disjoint subsets of the system.

3. The *a posteriori* repertoire

Suppose we have a subset S of a discrete system X . We detail the computation of the *a posteriori* repertoire $p(S_0 \rightarrow s_1)$ specified when S transitions to s_1 . The approach is completely general: it applies to the entire system X , or any part thereof.

To specify the *a posteriori* repertoire we need two pieces of data: the *a priori* repertoire of perturbations, and a description of the mechanism by which S responds to perturbations. The *a priori* repertoire is given by the maximum entropy distribution on the set of outputs of the elements. The mechanism is the causal architecture of the system; we suppose it to be encoded in the form of a conditional probability distribution: $p^{sys}(x_1|do(x_0))$, the probability the system outputs state x_1 given it is perturbed with x_0 , where the $do(-)$ operator is introduced following [2]. Since we are interested in the causal interactions that are accounted for internally by S , we ignore the responses of elements outside of S , and only consider the conditional distribution $p^{sys}(s_1|do(s_0))$. Perturbations introduced outside of S can affect the response by S , and are not accounted for internally, so we average (marginalize) over them. Each connection afferent to S is averaged over using the maximum entropy distribution. After averaging out extrinsic noise, the mechanism of an element n^k in S is given by:

$$p^{mech}(n_1^k|do(s_0)) = \sum_{e_0} p^{mech}(n_1^k|do(s_0, e_0)) \cdot p^{max}(e_0),$$

where the summation is over afferent connections from the environment $E := X \setminus S$ into S . The environment acts as a source of extrinsic noise that loosens the mechanism within S . The mechanism of the system, $p^{sys}(s_1|do(s_0))$ is then the combined mechanism of the elements within S . Imposing maximum ignorance on external inputs requires perturbing connections rather than elements, since connections from the environment to S are a property of S , and the external elements themselves are not.

Note that the alternate treatment of extrinsic noise, perturbing elements rather than wires, does not qualitatively affect the results presented in the text.

The *a posteriori* repertoire can then be captured via Bayes' rule:

$$p(s_0 \rightarrow s_1) := \frac{p^{sys}(s_1|do(s_0)) \cdot p^{max}(s_0)}{p(s_1)}$$

where $p(s_1) = \sum_{s_0} p^{sys}(s_1|do(s_0)) \cdot p^{max}(s_0)$. Bayes' rule provides a formal method for keeping track of the (probabilistic) effects of *a priori* perturbations introduced into the system, and for converting the resulting data into the *a posteriori* repertoire.

We consider two extreme cases. First suppose S is completely deterministic. It follows that

$$p(s_0 \rightarrow s_1) = \begin{cases} \frac{1}{m} & \text{if } p(s_1|do(s_0)) = 1 \\ 0 & \text{else} \end{cases}$$

where m is the number of perturbations in the *a priori* repertoire that cause s_1 . The *a posteriori* repertoire specified by S is simply all the perturbations that result in s_1 , weighted equally. From S 's point of view the

various perturbations in the *a posteriori* repertoire are indistinguishable. Note that the *a posteriori* does not specify the probability of s_0 leading to s_1 (which is 1 or 0 in this case), but rather the probability that s_0 led to s_1 .

Second, suppose S operates completely randomly: each outcome is equally likely regardless of prior inputs. This may be a result of the mechanism implemented by the elements in S , or due to external interactions overwhelming S . Either way, S does nothing to rule out alternatives by entering s_1 , and the *a posteriori* repertoire is $p(S_0 \rightarrow s_1) = p^{max}(S_0)$.

In an isolated deterministic system, Bayes' rule simply keeps track of which perturbations lead to s_1 , and renormalizes so that the probabilities sum to one as required. In a probabilistic system Bayes' rule keeps track of the additional data given by the likelihoods of the perturbations given the response, and weights the perturbations according to their likelihoods.

4. No integrated information is generated by a disconnected system

Suppose X consists of two disjoint subsets A and B . We show that $ei(X_0 \rightarrow x_1/\mathcal{P}) = 0$ for the partition $\mathcal{P} = \{A, B\}$, regardless of the state x_1 . Recall that relative entropy is zero if and only if the two repertoires under comparison are identical, so it suffices to show that the *a posteriori* repertoire specified by the parts coincides with that specified by the system.

If A and B are disjoint then the conditional probability $p(x_1|x_0) = p(a_1, b_1|a_0, b_0)$ splits into a product: $p(x_1|x_0) = p(a_1|a_0) \cdot p(b_1|b_0)$. It follows that the *a posteriori* repertoire also decomposes into a product, and so the combined *a posteriori* repertoires of the parts coincides with the *a posteriori* repertoire of the system.

5. The expectation of effective information

In the course of computing the *a posteriori* repertoire given state s_1 we computed the probability of the system entering s_1 given its causal architecture and the *a priori* repertoire of perturbations. It follows that we can compute the distribution $p(S_1)$ of possible outcomes at $t = 1$. We use $p(S_1)$ to compute the expected value of effective information.

First observe that effective information decomposes as a sum of simpler terms:

$$\begin{aligned} ei(S_0 \rightarrow s_1/\mathcal{P}) &= H \left[p(S_0 \rightarrow s_1) \left\| \prod_k p(M_0^k \rightarrow \mu_1^k) \right\| \right] \\ &= \sum_k H [p(M_0^k \rightarrow s_0) \parallel p(M_0^k \rightarrow \mu_1^k)] + H \left[p(S_0 \rightarrow s_1) \left\| \prod_k p(M_0^k \rightarrow s_1) \right\| \right]. \end{aligned} \quad (S1)$$

The expectation of $H [p(M_0^k \rightarrow s_1) \parallel p(M_0^k \rightarrow \mu_1^k)]$ is

$$\sum_{s_1} p(s_1) H [p(M_0^k \rightarrow s_1) \parallel p(M_0^k \rightarrow \mu_1^k)] = H (p(M_0^k \rightarrow M_1^k)) - H (p(M_0^k \rightarrow S_1))$$

and the expectation of $H [p(S_0 \rightarrow s_1) \parallel \prod_k p(M_0^k \rightarrow s_1)]$ is

$$\sum_{s_1} p(s_1) \cdot H \left[p(S_0 \rightarrow s_1) \left\| \prod_k p(M_0^k \rightarrow s_1) \right\| \right] = \sum_k H (p(M_0^k \rightarrow S_1)) - H (p(S_0 \rightarrow S_1))$$

In both formulae we are now computing entropy rather than relative entropy. Define the expected effective information generated by a subset S of X as

$$EI(S_0 \rightarrow S_1) = H (p^{max}(S_0)) - H (p(S_0 \rightarrow S_1)). \quad (S2)$$

It follows that the expectation of effective information across a partition is

$$\begin{aligned} EI(S_0 \rightarrow S_1/\mathcal{P}) &:= E_{S_1} [ei(S_0 \rightarrow s_1/\mathcal{P})] \\ &= EI(S_0 \rightarrow S_1) - \sum_k [EI(M_0^k \rightarrow M_1^k)]. \end{aligned}$$

Thus the expected effective information generated across a partition is given by the expected effective information generated by the system minus (over and above) the expected effective information generated by the parts.

Finally, define *expected* integrated information, $\Phi(S_1)$, as expected effective information across the partition that generates the lowest quantity of normalized expected effective information:

$$\Phi(S_1) = EI(S_0 \rightarrow S_1/\mathcal{P}^{MIP}). \quad (\text{S3})$$

The expected integrated information provides a useful characterization of the capacity of a discrete system.

6. Bounds on $\Phi(X_1)$

Computing effective information for every partition of a set S is computationally intensive: the number of partitions grows faster than e^n as a function of the number of elements n in the set. We show that restricting to bipartitions gives a lower bound on the expected integrated information $\Phi(X_1)$. Suppose we have a subset S of X and bipartition $\mathcal{P} = \{A, B\}$ of S . Notice that $EI(S_0 \rightarrow S_1/\mathcal{P}) \geq 0$ since the expression is an expectation of relative entropies, all of which are non-negative. It follows that

$$EI(A_0 \rightarrow A_1) + EI(B_0 \rightarrow B_1) \leq EI(S_0 \rightarrow S_1) \quad (\text{S4})$$

for any bipartition of S . This can be interpreted as saying that the whole is always (on average) at least the sum of its parts.

For any partition $\mathcal{P}$ we can form a refinement $\mathcal{P}'$ by subdividing one or more of the parts of $\mathcal{P}$. Thus a bipartition can be refined to form a tri-partition, and so forth. At the bottom we find the partition into the individual elements, and below that we place the total partition, where there are no parts and the system is taken as a whole. Equation (S4) has the following consequence:

$$EI(X_0 \rightarrow X_1/\mathcal{P}) \leq EI(X_0 \rightarrow X_1/\mathcal{P}') \text{ if } \mathcal{P}' \text{ is a refinement of } \mathcal{P}.$$

It follows that effective information for the minimum information bipartition provides a lower bound for $\Phi(X_1)$ and that effective information for the total partition provides an upper bound. An additional upper bound – for systems consisting of elements with identical outputs – is given by twice the effective information on the minimum information mid-bipartition, as can be seen by considering the normalizing constant.

Homogeneous systems. We show that expected effective information in a homogeneous system of binary elements is at most one bit: $EI(X_0 \rightarrow X_1) \leq 1$ bit. Expanding, we see that $EI(X_0 \rightarrow X_1) = H(p(X_1)) - H(p(X_1|X_0))$. The first term quantifies the repertoire of possible outputs of the system. However for a homogeneous system the elements act monolithically regardless of the input: they all either fire or all silent. Thus $H(X_1) \leq 1$ bits. It follows from the nature of the normalization for different partitions that expected integrated information is at most one bit.

Lattices. Expected effective information across the minimum bipartition in an $n \times n$ lattice of binary elements is at most $2n$ bits. There are $2n$ elements that lie on the boundary of a vertical or horizontal bipartition on an $n \times n$ lattice. These elements are the only elements that are affected by interactions between the parts, and they can at most generate 1 bit of information on average. If we assume the lattice wraps around to form a torus, as in the Game of Life grid in Figure 15, there are two cuts across which parts can interact and the upper bound increases to $4n$.

7. Relationship to effective information for stationary systems

In previous work [3, 4] the notions of effective information and integrated information were introduced for stationary multivariate Gaussian random process. Fix a bipartition $\mathcal{P} = \{A, B\}$. The effective information $EI(A \rightarrow B) = H(A^{max}) + H(B) - H(A^{max}, B)$ across the bipartition captures the *capacity* of the subsets for integrated information. Integrated information of the system was then defined as the capacity across the minimum information bipartition; we denote it by $\Phi_{stat}^{pert}(X) = EI(A \rightleftharpoons B^{MIB})$.

In this paper we generalize effective information as a time and state dependent measure. The measure quantifies the effective information generated by the process of ruling out a collection of alternatives from the *a priori* repertoire. It cannot be directly connected to the previous notion since they apply in different contexts, however, we investigate how expected integrated information can be adapted to stationary systems since *expected* effective information can be interpreted as a capacity: the capacity for interactions between a collection of parts.

Recall that the expected value of effective information, Eq (S2), is given by $EI(A_0 \rightarrow A_1) = H(p^{max}(A_0)) - H(p(A_0 \rightarrow A_1))$, and expected effective information over an arbitrary partition is $EI(X_0 \rightarrow X_1/\mathcal{P}) = H(p(A_0 \rightarrow A_1)) + H(p(B_0 \rightarrow B_1)) - H(p(X_0 \rightarrow X_1))$. The most natural way to apply this to a stationary (timeless) system is to drop the time dependence and rewrite it as

$$MI(X/\mathcal{P}) = MI(A; B) = H(A) + H(B) - H(X),$$

the mutual information over the partition. Finding the minimum information partition, suitably normalized, and computing mutual information across the *MIP* recovers the *perturbation free* modification of integrated information discussed in [3]:

$$\Phi_{stat}^{corr}(X) = MI(X/\mathcal{P}^{MIP}).$$

It provides a measure of the minimal observed dependencies in a stationary system, and could be applied, for example, to determine the natural components of a stationary system.

Perturbing elements in a stationary system allows one to observe the effects one part can have on the rest of the system, but not the effects the entire system can have on itself. The perturbation free measure above detects correlations between subsets of the system, but does not detect causal interactions. This paper avoids these problems by considering non-stationary systems so that perturbations are introduced in one time step, and the response is measured in the next.

Finally, note that integration (not to be confused with integrated information) as introduced in [5] is given by $MI(X/\{X_k\}_{k=1}^N) = \sum_{k=1}^N H(X_k) - H(X)$.

8. Comparison with other information measures on discrete systems, and discussion of technical aspects of ϕ

It is worth briefly discussing the relation between the present work, and the notions of stochastic interaction, transfer entropy and information flow.

A measure of stochastic interactions within a discrete system was introduced in [6], motivated by the desire to extend the infomax principle of [7] to dynamical processes. In their work they found that optimizing stochastic interaction results in systems with almost deterministic global behavior, and long cycles with relatively few branch points. Further, the individual elements in the optimized systems were extremely poor predictors of the system's future behavior. Thus the optimized systems possess an intriguing mix of local flexibility and global rigidity, suggesting they may generate high ϕ .

Stochastic interaction is defined as follows. A discrete system X is given, with an initial probability distribution p on the outputs of the elements (in examples p is a stationary distribution) and causal mechanism M . The system is partitioned into its individual elements: $\mathcal{P} = \{\{n^k\}_{k=1}^N\}$, and for each element n^k the distribution p is

projected to its marginal distribution p^k . The causal mechanism (Markov kernel) M^k of each part is computed in a manner similar to that in this paper. Stochastic interaction is then

$$I(p, M) = \sum_k H(p^k, M^k) - H(p, M),$$

where $H(p, M) = \sum_{x_0} H(M(X_1|x_0))$. By comparison, *expected* effective information over and above the partition into individual elements is

$$EI(X_0 \rightarrow X_1 / \{\{n^k\}_{k=1}^N\}) = \sum_k H(p(N_0^k \rightarrow N_1^k)) - H(p(X_0 \rightarrow X_1)).$$

Thus, in a special case stochastic interaction (with the maximum entropy distribution) can be seen to be similar to a special case of expected effective information (across a particular partition, i.e. the partition into individual elements). The underlying motivations however are quite different. Stochastic interaction measures the total statistical dependence introduced into the system by the action of the Markov kernel and can be thought of as a dynamic implementation of the notion of integration in [5]; whereas ϕ is a measure of the repertoire of states specified by the system as a whole.

Contrasting the two measures is a useful exercise, since they seek to answer different questions using similar information-theoretic tools. We emphasize the differences, not to criticize the notion of stochastic interaction, but rather to provide a stage for motivating and summarizing certain technical aspects of the definition of ϕ , complementing the phenomenological presentation in the main text.

- (a) *Stationary versus maximum entropy distributions.* The stationary distribution(s) of a Markov process are for many purposes the most natural to use in an analysis of its dynamics, since it describes the equilibrium behavior. ϕ does not capture or characterize equilibrium behavior, but is rather concerned with how a system reduces *a priori* ignorance. The maximum entropy distribution captures the notion of maximal ignorance; which is then reduced by the mechanism. It also has the important consequence that ϕ is bounded, which is not true in general of relative entropy.
- (b) *Averaging over states versus considering a particular state.* Since we are interested in the information generated by the system, we attend to the particular state it enters, rather than take an average of states.
- (c) *Looking forward versus looking backwards.* Stochastic interaction quantifies how the Markov kernel transforms a distribution at $t = 0$ to a distribution at $t = 1$, “counting” how many different outcomes the system generates compared to its elements; i.e. by looking forward at where the system could go. In contrast, effective information quantifies the information generated by the system transitioning from x_0 to x_1 using the *a priori* and *a posteriori* repertoires; i.e. by looking back at the possible prior perturbations through the lens of the mechanism. Looking forward quantifies the number of outcomes the system can produce in general, without attending to the actual state the system is in; whereas looking back asks how much information the system generates upon entering its actual state. Consider, for example, a six-sided die. We take the position that the die does not generate information in and of itself (i.e. because it is six-sided); but rather that it must be thrown, and generates information in the act of landing on a particular face, ruling out the alternatives.

Returning to the analysis of discrete systems, as an external observer all we see is the first state followed by the second (note that x_0 may in fact be the same as x_1 : the mechanism may cause the system to stay in the same state). However, the system, we have to assume, has certain undeniable properties:

- (i) at $t = 1$ it is in a particular state;
- (ii) it has a mechanism, which admits a description as an input-output table; and
- (iii) it has a repertoire of possible inputs; about which the system has no *a priori* information.

It follows that there is a certain amount of effective information associated with entering any particular state. It is fruitful to visualize the mechanism as a filtering device. Each input in the *a priori* repertoire leads to a particular output, so the mechanism acts by grouping inputs into boxes labeled by outputs (possibly

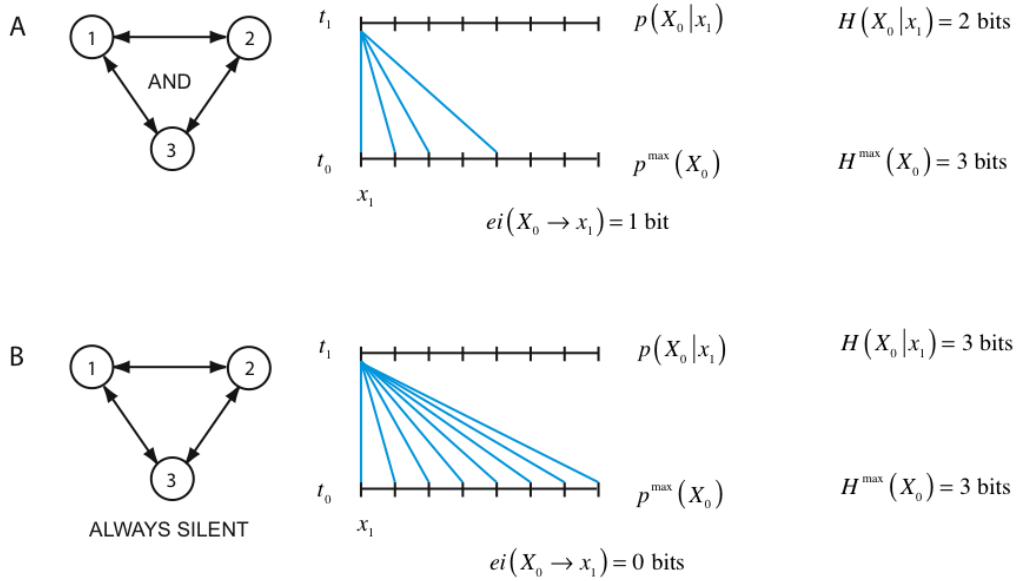


Figure 1: Two systems in a steady state with no elements firing. (A): a system of three *AND*-gates. The system generates 1 bit of effective information since at each time step the elements *choose* not to fire after checking their inputs for spikes. (B): a system which is silent no matter the prior state. No effective information is generated.

probabilistically). It is the action of the mechanism, taking inputs and determining the corresponding outputs, that generates information. The particular state the system is (or was) in is not informative in and of itself. Thus we are not interested in where the system was or even where it is per se, but rather in how the mechanism reduces uncertainty by entering the current state. The answer is given by the rules implemented by the elements, which we unfold by using the *a priori* repertoire to determine how the rules act on different possible states. Perturbing with all possible states ensures we exhaustively uncover the processing of the mechanism. By looking back at the prior perturbations through the lens of the mechanism, from the vantage point of the current state, we see how the *a priori* repertoire is contracted into the *a posteriori* repertoire by the processing of the system. Contrasting the two repertoires captures the “differences that make a difference” to the system.

As an example, consider Figure SI-1. In panel A the system is in a steady state, with no elements firing. Nevertheless, it generates 1 bit of integrated information each time it does nothing (the minimum partition in this case is the total partition). The mechanism causes the system to actively choose to remain silent. From the vantage point of the current state, the four *a posteriori* states that lead to state x_1 – which the system cannot distinguish among – are specified against the background of the eight possible *a priori* states, through the lens of the system’s mechanism. The system in panel B is also in a fixed point, with no elements firing. However, in contrast to panel A, this is not as a result of causal interactions, but rather a result of the elements ignoring their inputs and remaining silent no matter what. For this system we find that integrated information is zero bits, although to the external observer its behavior is identical to that in panel A.

- (d) *Difference of entropies versus directly comparing repertoires.* Stochastic interaction is the difference of the entropy generated by the elements and the entropy generated by the whole. By contrast, and crucially for our purposes, effective information directly compares the repertoires of the parts and the whole, without first converting each into a number, and then comparing the two numbers. We quantify the information generated by the system, over and above that generated by a collection of parts, by computing the entropy of the *a*

a posteriori repertoire of the system relative to the combined *a posteriori* repertoires of the parts. A conceivable alternative would be to subtract the entropy of the system's *a posteriori* repertoire from the sum of the entropies of the *a posteriori* repertoires of the parts. This would provide a measure that compares numerically the reduction in uncertainty performed by the parts with that done by the whole. Note however that we wish to capture how the processing of the whole differs from that of the parts: the *a posteriori* repertoire serves as a means to expose the action of the mechanism, rather than an end to be quantified in itself. It is therefore necessary to determine, for each perturbation taken separately, how the processing of the system differs from the combined action of the parts. This ensures that integrated information, as defined in the text, is zero if and only if the parts and the whole specify *identical* repertoires, which is not necessarily the case if we simply compare the sizes of the repertoires. Thus, rather than simply compare the sizes of the repertoires as seen through the lenses of the whole and the parts, we directly contrast the distributions, perturbation by perturbation. In this way, computing relative entropy, we directly compare the processing of the system as a whole to the system taken as a collection of parts.

- (e) *Partitions.* Stochastic interaction is computed using the partition into individual elements. Suppose we have a system of 4 elements, decomposing into 2 separate couples with internal connections, but no connections between the couples. Stochastic interaction will be nonzero, which makes sense, since there are indeed interactions within the couples. However any measure of integration must be zero on a disjoint system, since a disjoint system is not a single entity. Thus ϕ is computed by comparing effective information across all partitions.

We complete the discussion by considering the notions of transfer entropy (TE) and information flow (IF). Transfer entropy [8] can be written (for special case $l = k = 1$) as

$$T_{A \rightarrow B} = \sum_{a_0, b_0} p(a_0, b_0) \cdot H[p(B_1|a_0, b_0) || p(B_1|b_0)].$$

It is an average of relative entropies over some initial probability distribution. It captures how much information flows from A to B by measuring how much variability in B is explained by A . Information flow [9, 10] is a similar measure capturing the flow of information between two subsets of a discrete system using causal interventions.

We do not compare effective information with these measures in detail. Instead it is illuminating ask the question: what prevents transfer entropy and information flow from forming suitable foundations for measuring integrated information? Either measure could be computed across bipartitions, using $T_{A \rightarrow B} + T_{B \rightarrow A}$ for example. Comparing across all bipartitions using a normalization, would then result in measures superficially similar to ϕ . This sort of approach arguably provides the most straightforward way to adapt effective information from stationary to non-stationary systems, so it is worth explaining why we did not follow this route.

To avoid repeating the presentation in the text, and the discussion of stochastic information above, we restrict attention to a single aspect of our reasoning. Integrated information is motivated by a desire to understand how a discrete system acts (generates information) as a single entity. We directly compare the processing performed by the system as a whole with the processing performed by the parts – considered as wholes in their own right – to quantify the interactions in the system that *cannot be reduced* to the actions of independent subsystems. In contrast, TE and IF decompose a system into parts from the start: they quantify the flow of information (suitably defined) from one part, A , of a system to another, B . Thus, TE and IF take a fundamentally local perspective on the system that precludes understanding it as a single entity.

A simple example illustrates the point. Consider the parity system of Figure 14. The system behaves as a single giant element, and generates 1 bit of integrated information. However if we try to measure integrated information using TE or IF , splitting the system into two parts, we find that each part is capable of making one bit's worth of difference to the other, resulting in two bits of integrated information. By focusing on the effects of the parts on each other, rather than that of the system on itself (relative to the parts), we obtain an incorrect measurement. More generally, suppose we calculate the TE or IF between all the subsets of a system. It is not clear how this data can be combined to generate a single coherent picture of the processing performed by the system. Our

solution is to emphasize the action of the whole over the parts, rather than the action of the parts on each other.

Note that we could not do this without introducing time as a variable into the systems. Thus previous work [3] necessarily considered the actions of parts on each other, rather than the whole relative to the parts, since (as far as we know) this is the only way to analyze causal interactions in stationary systems.

9. Applications of ϕ and approximations

We briefly discuss possible approaches to applying ϕ to neural systems. There are at least two major hurdles. The most serious is that the causal architecture of a system must be known before ϕ can be measured. A further daunting problem is dealing with the combinatorial explosion in the number of partitions and states that have to be considered in large systems.

The computational burden can be reduced in a number of ways. First, the discussion of expected effective information and Φ above shows that on average it is sufficient to restrict attention to bipartitions. Second, it is often the case that most bipartitions can be eliminated as possibilities out of hand in a system that is well-studied and understood, recall the grid for example. Third, the computation of the *a posteriori* repertoires can potentially be short-circuited. For example, perturbations that involve an abnormally high or low firing rate can be dismissed. Finally, the system in question can be considered over longer time frames and approximated as a stationary system, so that the mutual information can be used instead of effective information, see [4].

Although these approximations may prove useful and interesting, a more basic point should be made. According to the theory, a neural (or any other) system is a single entity insofar as it generates a large amount of integrated information. A practical way of addressing this issue is to consider the two notions (integration and information) separately. Approximate measures of integration can be developed; for example based on functional clustering [11], or using the mutual information approximation to integrated information, or possibly by adapting transfer entropy or information flow. Although none of these measures fully captures the notion of integration, they provide useful tools in the absence of a detailed understanding of the causal mechanisms underlying mammalian brains. Once integration has been approximately established, further measures such as coherence dimension or stochastic interaction could be applied to quantify the degrees of freedom available to the system considered as a single entity, providing an approximate measure of the integrated information generated.

10. Integrated information and the choice of elements and variables

We have seen in the section on complexes that it is important to properly choose the elementary components of a system. An instructive example is given by the deterministic Hopfield-type network of [12]. It has connection weights from the j^{th} element to the i^{th} element given by 2^j , so that network activity is updated according to

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

where $x_j(t)$ describes the state of the j^{th} neuron at time t . Computing integrated information they find that $\Phi = n$ bits. They then go on to observe that the network can be rewritten in the form

$$v_i(t) = f\left(\sum_{j=1}^n 2^j v_j(t-1)\right),$$

where $v_i(t) = f(x_i(t))$ is a binary variable describing the *output* of the i^{th} neuron, in which case $\Phi = 2$ bits. The claim is that there is a family of systems with simple dynamics and interactions between elements, and with arbitrarily high Φ . Moreover the discrepancy between the Φ values in the two realizations of the system would show that Φ is sensitive to a change of variables since the two systems “produce identical behavior”.

The example is a discrete system, though their adaptation of Φ and effective information from [3] differs from the

one provided here. Nevertheless, the error in the example rests on the treatment of causal interactions, rather than in the implementation of Φ .

Let us first consider the claim that it is possible to construct simple systems with arbitrarily high values of Φ . The possible states of a single neuron are never explicitly described in [12], except to say the states are integer valued. By considering all possible 2^n initial conditions for the network, we see that each neuron has at least $N = 2^n$ internal states. In particular each neuron stores at least n bits. Since $\Phi = n$ it is apparent that the network is integrating no more information than the individual neurons. The difficulty involved in constructing a system with high Φ has been shifted onto the shoulders of the neurons: it is not clear how to build neurons that are integrated objects with repertoires of 2^n for arbitrarily large n . This relates to the discussion of Figure 7AB above. If we take the units to be single objects then it is easy to construct systems with high Φ by coupling them. The problem is that the units are not integrated.

Second, consider the claim that Φ depends on the choice of variables. Is the natural choice the 2^n internal states or the 2 outputs of the elements? Since Φ measures causal interactions internal states are only relevant to the extent that they contribute to outputs; beyond that they are not used by the model and are in fact somewhat artificial. Once we – or rather the elements – have identified whether the internal states are positive or negative they have no further causal consequences. The difference in Φ obtained in the two computations is not a consequence of changing variables, but rather a result of measuring things that are not effectively there. Φ is 2 bits regardless of the number of elements and connections, reflecting the simple causal architecture.

11. Integrated information for continuous systems

The theory developed in the paper assumes that elements have a finite repertoire of outputs, and that time passes in discrete instants. We do not develop the theory for continuous systems; indeed, it is unclear whether continuous variables exist in nature. However, it is suggested in [12] that ϕ can be infinite and depends on the choice of scale for systems with continuous variables, so we briefly discuss these issues.

Relative entropy for continuous random variables (which have a probability density function) is

$$H[p \parallel q] = \int_X \log_2 \left(\frac{p}{q} \right) dp,$$

where q is the prior and p the posterior. Relative entropy, applied to measure effective information, is easily seen to be finite (briefly, the support of p is contained in that of q since p is specified using more data than q). Relative entropy is also invariant under rescaling: if we change the units of measurement, rescaling from $X = [0, T]$ to $X = [0, \alpha \cdot T]$ for example, then the scaling factor affects both p and q and cancels out. Thus systems with continuous variables pose no theoretical difficulties.

A related question is whether continuous variables provide a cheap method for constructing systems with high ϕ . It is possible to imagine systems (whether discrete or continuous) of coupled components with large repertoires that have high ϕ . As a practical matter, the question is always how such a system could be implemented. The world appears, at the smallest scales, to be comprised of discrete components with limited repertoires of available states. Thus any object X with a large repertoire must contain a correspondingly large number of components. However, as discussed in the section on Complexes, see Figure 7, in that case two possibilities obtain. If the elements within an object are independent then ϕ will be zero. On the other hand, to the extent the elements within an object are coupled, the *a posteriori* repertoire of the object will shrink, so effective information between objects is reduced. Finally, it is also important to note that, since ϕ measures causal interactions between interacting elements, one should be careful to identify variables that reflect causal states within a system. Specifically, macroscopic variables such as phase or position, which are typically identified by an external observer, are only relevant insofar as they correspond to intrinsic properties of the elements of a complex that make a difference to their interactions.

12. Boolean functions in scrambled counting network

Each element in the scrambled counting network implements a Boolean function, which we detail. For each element we list the 8 firing patterns that cause it to remain silent. It will fire when it receives any of the remaining 8 inputs.

Element 1	0000	0001	0101	0110	1000	1011	1110	1111
Element 2	0000	0001	0011	0111	1010	1011	1100	1111
Element 3	0000	0001	0101	0110	1001	1010	1100	1101
Element 4	0001	0010	0110	0111	1001	1011	1100	1110

13. Connection matrix of network in Figure 18

Row 1 shows efferent connections from element 1 to elements 1,2,3 and so forth. Blank entries signal there is no connection.

	.875	-.375			-.875	.75	
.5		-.375	1	.875	-.125	.5	.5
-.25	-.875		.125	.125		.5	
-.75	-.375			-.125		-.75	
-.625		-.875			-1	1	
	-.625	.75		-.375		-.5	
-.875		.5		-.125	-.5		-.25
-.75	-.25		-.375		.25	-.875	

References

1. Cover T, Thomas J (2006) Elements of information theory. John Wiley & Sons.
2. Pearl J (2000) Causality: models, reasoning and inference. Cambridge University Press.
3. Tononi G, Sporns O (2003) Measuring information integration. BMC Neuroscience 4:31.
4. Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience .
5. Tononi G, Sporns O, Edelman G (1994) A measure for brain complexity: Relating functional segregation and integration in the nervous system. Proc Nat Acad Sci 91:5033–5037.
6. Ay N, Wennekers T (2003) Dynamical properties of strongly interacting markov chains. Neural Netw 16:1483–1497.
7. Linsker R (1986) From basic network principles to neural architecture. Proc Nat Acad Sci 83:7508–7512.
8. Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85:461–464.
9. Ay N, Krakauer D (125) Geometric robustness theory and biological networks. Theory Biosci .
10. Ay N, Polani D (2008) Information flows in causal networks. Adv Complex Syst 11:17–41.
11. Tononi G, McIntosh A, Russel D, Edelman G (1998) Functional clustering: Identifying strongly interactive brain regions in neuroimaging data. Neuroimage 7:133–149.
12. Seth A, Izhikevich E, Reeke G, Edelman G (2006) Theories and measures of consciousness: an extended framework. Proc Nat Acad Sci 103:10799–10804.