Nonlinear Dynamics and Neuroscience

Izvestiya Vysshikh Uchebnykh Zavedeniy. Applied Nonlinear Dynamics. 2023;31(2)

Review DOI: 10.18500/0869-6632-003033

Integrated information theory and its application for analysis of brain neuron activity

I. A. Nazhestkin[1] ✉, *O. E. Svarnik*[1,2]

[1]Moscow Institute of Physics and Technology (National Research University), Russia
[2]Institute of Psychology of Russian Academy of Sciences, Moscow, Russia
E-mail: ✉nazhestkin@phystech.edu, olgasvarnik@gmail.com

Received 30.10.2022, accepted 18.02.2023, available online 2.03.2023, published 31.03.2023

Abstract. *Purpose* of this review is to consider the possibility to apply the integrated information theory to investigate the brain neural activity. Earlier was shown that the integrated information amount Φ quantifies a degree of a dynamic complexity of a system and able to predict a level of its success defined by classic observable benchmarks. For this reason, a question arises about the application of the integrated information theory to analyse changes in brain spiking activity due the acquisition of new experience. *Conclusion.* The bases of the integrated information theory and its possible application in neurobiology to investigate the process of new experience acquisition were reviewed. It was shown that the amount of integrated information Φ is a metric which is able to quantify the dynamic complexity of brain neural networks increasing when the new experience is acquired. Methods, enabling the practical calculation of Φ for spiking data, were proposed.

Keywords: brain, information, learning, integrated information theory, complexity.

Acknowledgements. This work was supported by Russian Foundation for Basic Research, grant No. 20-013-00851.

For citation: Nazhestkin IA, Svarnik OE. Integrated information theory and its application for analysis of brain neuron activity. Izvestiya VUZ. Applied Nonlinear Dynamics. 2023;31(2):180–201. DOI: 10.18500/0869-6632-003033

Fundamentals of integrated information theory

Integrated Information Theory (II) and its basic concept — integrated information coefficient Φ — were originally proposed to assess the level of consciousness in the brain [1–4]. With the help of IIT, an attempt was made to explain what are the features of neural activity underlying the phenomenon of conscious subjective sensations, why they appear precisely due to the work of the brain and whether they arise in other complex systems. J. Tononi suggested that consciousness

arises only in those systems where information is integrated. The coefficient Φ evaluates two abilities of a complex system that are critical for its operation — segregation of information and its integration. Segregation means that the system should be able to be in as many states as possible in order to generate as much information as possible. At the same time, it is important that the components of the system are highly specialized, that is, they are associated with coding, for example, only one type of stimuli. Integration consists in the fact that such components of the system must be interconnected. This is critically important for the operation of the system as a whole. Small values of Φ indicate either weak generation of specialized information by system components, or that there is a place in the system with weak interconnection of components. Both have a negative effect on its performance in tasks observed from the outside using classical criteria.

To better understand why both segregation and integration are important, Tony and colleagues cite an analogy with the photosensitive matrix of a camera. Each of, for example, a million pixels can be in one of two states (light falls on it or does not fall on it), and thus the matrix is able to encode $2^{1,000,000}$ possible states. This means that the ability of the matrix to differentiate information is very high — it is able to distinguish a very large number of incoming images. However, there are no interdependencies between the pixel states, they are completely independent and there are no connections between them. Therefore, no conclusions can be drawn from the encoded information, no processing of the received image in such a system is impossible. Thus, the system is only capable of storing information, but not using it for processing and obtaining conclusions.

Using IIT, results were obtained that are in good agreement with the already available experimental data on the small contribution of the cerebellum and the large contribution of the cerebral cortex to the processes of consciousness [5], and differences in the predicted level of consciousness in the states of wakefulness, sleep [6], anesthesia [7] and for various brain injuries [8]. This can be useful in medicine to determine whether the patient is conscious or not [9, 10]. However, in this application, the theory has been seriously criticized. Firstly, she attributed a high level of consciousness to systems that do not have consciousness based on common sense, for example, computers. Secondly, consciousness does not depend on the observer, unlike information, therefore, based on information, it is incorrect to declare the presence of consciousness. Finally, a set of discrete states of the system is used to calculate the indicator Φ, which, although sufficient in most cases, leads to information loss — for example, for the same neurons whose local field potentials encode a large amount of information [11, 12]. Therefore, the question of describing the level of consciousness by this theory remains open.

At the same time, it was shown that the integrated information coefficient Φ well describes the performance of other complex systems consisting of a certain number of interacting elements [13, 14]. According to the researchers, Φ shows the internal state of complex self-organizing systems consisting of individual elements in which order arises at a certain level of complexity. Φ describes the internal structure leading to a certain behavior of the system, to the complexity of the states arising in it ("What the system is"), and does not describe the observed states of the system from the outside ("What the system does") [15, 16]. So, in the work [14], the coefficient of integrated information was calculated on two data sets: for teams of people performing tasks directly related to team interaction, and for a group of editors of articles in Wikipedia. In the first case, 68 groups of people, out of 4 people each, solved various tasks, for example, answered intellectual questions by brainstorming or typed text in a common online editor. Common to all tasks was that team members had to interact with each other to successfully complete the

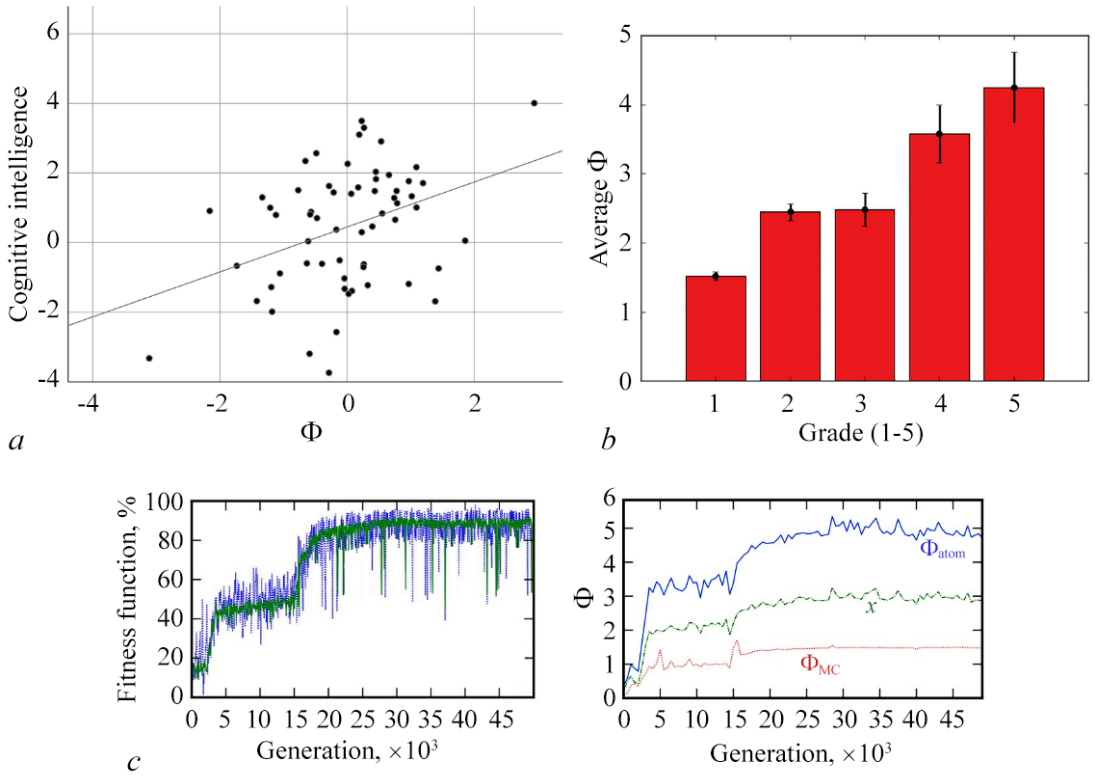

Fig. 1. Examples of the integrated information theory application for analysis of the action of different complex systems. *a* — Correlation between an integrated information coefficient and a collective intelligence for human groups. *b* — Relation between groups of Wikipedia editors and manuscript grades (1–5). Changed from [14]. *c* — Application of an integrated information coefficient calculated with different methods to quantify the behavior of a virtual entity in a maze. Left: learning process; a normalized fitness function is shown. Right: evolution of integrated information coefficient calculated with three different ways. A correlation between Φ and a fitness function is observed. Changed from [13] (color online)

task. The nodes in the network were people who were assigned the state "1"at each moment of time, if the person was talking at that time, and "0"if he was silent. Thus, the coefficient Φ estimated the generation of information in a network consisting of people. As a result, the collective intelligence of the teams (the average number of points scored by the teams during the execution of all tasks) was statistically significantly correlated with the value of Φ calculated during the work of the teams (Fig. 1, *a*). The second data set examined the editing history of 999 of the most important Wikipedia articles. For each article, all the people who edited it were considered. The coefficient of integrated information Φ for such groups of people turned out to be statistically significantly related to the evaluation of their article issued by users on the site (on a five-point scale) (Fig. 1, *b*).

In [13], the coefficient of integrated information was calculated for virtual model organisms ("animats") trained in the skill of orientation in a maze. The organism had sensors, moving organs and internal intermediate elements for processing information, each of which was a node of the network having a state of "0"or "1". The network of these elements worked according to the probabilistic principle, with given probabilities of transition from state to state. The organism needed to pass from left to right a randomly generated maze with many vertical walls having one door in different places. The training of organisms was carried out using a genetic algorithm, which is based on the application of random changes (mutations) and random fusion

of generated systems (crossing) at each iteration (generation). At the same time, the transition probabilities and the structure of connections between elements changed from generation to generation. As a result, it turned out that the coefficient of integrated information calculated for network elements (not for transition probabilities!) it increased with the course of training. Similarly, the fitness function increased, showing the success of the organism, and depending on the number of passes through the doors and the distance to the last door (Fig. 1, *c*). Thus, the indicator Φ presumably evaluates the nature of the system's collaboration, the complexity of interactions between its elements, ensuring performance in the tasks performed by the system.

For the brain (in vivo and in simulations), the Φ coefficient also showed a relationship with brain development and the success of the behavior performed. In [17], a computational model of the brain in which 5.2 million neurons were simulated was trained according to the Hebb rule. Learning was characterized by an increase in the number of attractors (formed as a result of self-organization of stable patterns of neuronal activity, which, according to available data, encode certain stimuli [18, 19] and thus directly ensure the segregation of information). After the training, an increase in the indicator of integrated information was noticed. To calculate Φ by 5.2 million model neurons from the membrane potentials of neurons, signals recorded using electroencephallography (EEG) according to Coulomb's law [17] were simulated. In the work of J. Islera et al. [20] the coefficient of integrated information was calculated for EEG recorded in newborn children, depending on age (in days). It has been shown that the indicator Φ increases with age. Since active brain development occurs in childhood, this fact confirms the hypothesis that Φ describes the development of a complex system. In the work [21], when teaching animals spatial skill, the coefficient Φ for the neural activity of the hippocampus, calculated in various ways, had a positive correlation with the metric of learning success, depending on the environment — the number of rewards given to the animal. In [22], a similar result was shown with spatially aversive (fear-related) learning for the neural activity of two brain regions (hippocampus and amygdala) (Fig. 2, *a*). In addition, on the days when there was an increase in the number of rewards, but not in subsequent ones, a correlation was shown between Φ for hippocampal neurons

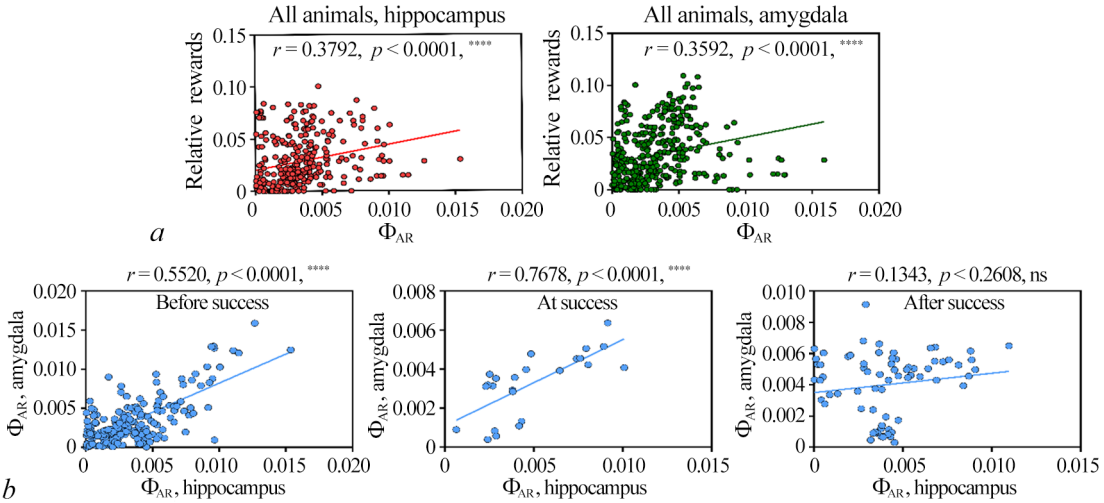

Fig. 2. Example of the integrated information theory application for quantification of the degree of learning success. *a* — correlation between Φ for hippocampus and amygdala neurons and a relative number of rewards (normalized by the learning session duration). *b* — correlation between Φ for hippocampus and amygdala neurons: left — before the day when the maximal number of rewards was achieved, middle — during a such day, right — after this day until the end of learning. Changed from [21, 22] (color online)

and Φ for amygdala neurons (Fig. 2. *b*).

These results show that segregation and integration of information, described by the coefficient of integrated information Φ, are interrelated with the success of the functioning of a complex system. Therefore, the application of the theory of integrated information to the analysis of neural activity of the brain seems to be a promising approach. The coefficient Φ is able to show internal changes in a complex system, therefore it will allow taking into account previously unknown patterns in neural activity patterns. This will allow us to more accurately assess the degree of involvement of certain neural groups in the learning process and the level of plastic changes occurring in the brain during the acquisition of skills.

Fundamentals of the processes of segregation and integration of information

The successful application of IIT for the analysis of various systems shows that segregation and integration of information are found in various self-organizing systems [13, 14]. The brain is no exception. Such processes directly follow from the network structure of the brain formed by neurons and their interactions. The network structure of the brain is represented by communities with close internal mutual connections. It was shown [23] that neural communities implement segregation of information, and the connections between them — integration. Let's look at these processes in more detail.

Back at the beginning of the XX century, using the simplest microscopy methods, it was shown that the brain is not homogeneous, but has a large number of pronounced anatomical areas [24, 25]. Later, clusters of neurons [26] were found inside the brain, the structure of which is determined by the anatomical structure of the brain [27]. With such a network arrangement, special subgroups of neurons are formed in it, which contain a large number of connections — significantly more than all other neurons. This subgroup is called the "Rich club". Such cells are found in each cluster and are responsible for communication and information exchange between clusters. As a result, it turns out that each cluster can contain cells that have connections with other clusters. Such cells are called hubs. The "Club of the Chosen"forms a powerful structure that provides connections between clusters, allowing information to quickly move from almost any cluster to any. It is shown that in comparison with a random network, such a structure allows to significantly reduce the length of the path between the regions [28], which, as will be shown later, is critically important. The formation of a "club of the chosen"is described by the Barabashi-Albert [29] network model, where connections at each node are formed with a probability proportional to the already existing number of connections of this node. The proposed generative model explains a similar structure — cells from the "club of the chosen ones"are more likely to form synaptic connections than the rest, therefore they are more likely to connect with cells from all other clusters, and not only with the nearest ones. In such networks, the global clustering coefficient C is high, defined as the average for all nodes ratio of the number of interconnected pairs of neighbors of nodes to the total number of pairs of neighbors of nodes. It shows how strongly the hub nodes with a large number of connections are expressed in the network.

Each cluster, in turn, is built internally according to certain rules, which are the direct opposite of the organization of hubs. A good specialization in performing a function requires a quick exchange of information within the community. To do this, it is necessary to minimize the length of the pathways between neurons in the community. And neurons in specialized regions of the brain do form such networks, which has been confirmed by various methods [30–33].

The structure of neural connections in such clusters is described using a model of the so-called "small-World networks". Initially, such networks were proposed to describe the graph of social connections between all people — it was revealed that on average only 6 acquaintances share two randomly taken people, and that people using only their acquaintances, only locally accessible people, independently find these paths [34]. This shows the high speed of information exchange in such networks. The emergence of a "tight world"network is described by the Watts-Strogatz [35] model: in a regular ring network with k, a random test is performed for each connection with a probability of a positive outcome p, and in the case of a positive outcome, this connection is redirected by one of the ends to another node randomly selected in such a way that there was no similar connection before redirection. At $p = 0$, a completely regular network is formed (which is obvious, since no connections in the original regular network will be modified), and at $p = 1$, the resulting network will be completely random. Thus, the network in each region of the brain is the result of a balance between randomness and regularity. The most important property of the "close world"networks is that they have an extremely low average path length between two elements, weakly depending on their number. According to a study by Watts and Strogatz, the average path length in such networks is determined as follows:

$$\langle l \rangle = \frac{\ln N}{\ln K},$$

where N is the number of elements in the network, and K is the number of connections of one element. If we take the number of neurons in any anatomical region $N = 1$ billion, and the average number of synapses in a neuron $K = 1000$, we get the average length of the connection $\langle l \rangle = 3$. Such a small length allows such networks to generate information faster. It can be shown that in such networks the measure of the efficiency of information exchange is much higher:

$$E(G) = \frac{1}{N(N-1)} \sum_{i,j \in G} \frac{1}{d_{i,j}},$$

where G is the network under consideration (graph), N is the number of nodes in the network, and d_{ij} is the length of the shortest path between nodes i and j in the network. Normalization by $N(N-1)$ is needed in order to take as a unit the efficiency of a network having all $N(N-1)/2$ possible connections [36, 37]. Artificial neural networks built on the basis of the "close world"model learn faster and make fewer mistakes [38].

A similar pattern exists at higher levels. There is every reason to assume that networks in the brain are multilevel, that is, nodes can be not only neurons, but also entire networks built on the same principle [39], that is, the brain can be described as a hypernet [40]. Artificial structures (reconstructed or created based on the application of graph theory), for example, voxels (three-dimensional pixels used for the analysis of functional magnetic resonance imaging, fMRI), can also be distinguished as nodes. The relationship between nodes in such studies is determined based on their activity by various statistical methods [41–44], after which close communities are distinguished using a special class of algorithms for detecting related populations (Communiy detection) [45, 46]. In such studies, voxels also showed a similar network structure, with close communities and hubs [47–50]. For some areas of the brain, the equivalence of the structure obtained by anatomical methods and methods of analyzing the activity of groups of neurons using the fMRI [51] method was shown. Thus, a similar structure is found at different levels. Perhaps the network structure at each level determines the structure of the next level — and so on, up to higher levels of networks, like social networks, directly accessible for observation.

Apparently, such networks arise evolutionarily under the influence of a complex environment. It is quite difficult to trace the evolution of a living brain network due to the great complexity of such networks, therefore computer modeling is used. In [52], a model of a multilayer neural network with connections changing according to a genetic algorithm was trained for several tasks, and the required task was often changed. At the same time, structural motives similar to the above-described motives in the networks of the living brain appeared in such a network. A specially defined modularity metric, showing how well communities can be identified in the network with a minimum of connections between them, grew with such training. A simpler linear model, in which the vector of results was obtained from the vector of environmental resources using a matrix, and the success metric was calculated as the modulus of the difference between the vector of results and a given ideal vector, showed a similar result — blocks of non-zero elements were found in the matrix against a general background consisting mainly of zeros. Here, the frequency of task changes turned out to be directly related to a similar modularity metric for the matrix [53].

Thus, in the brain, as a result of evolution in the environment, two directly opposite "standards"of network formation coexist — networks of a "close world"with a large number of connections between different pairs of elements and networks with nodes having a large number of connections. The former are characterized by a low clustering coefficient, high link density and a small average path length, and the latter — on the contrary, a high clustering coefficient, low link density and an increased average path length. The activity of nodes in each community strongly correlates with each other and weakly correlates with activity in other communities. These types of networking are directly related to differentiation (segregation) and integration of information [23].

Segregation of information means that the processing of highly specialized information of each specific type, the performance of each specific task is carried out in its own specialized modules. These modules are precisely the aforementioned communities that have close internal ties according to the "close world"model. In them, firstly, short connections between elements increase the speed of information generation, and, secondly, a group of a large number of neurons can be in a large number of states, and thus distinguish a large number of objects.

Integration of information — the presence of relationships, interdependencies between the states of nodes in different communities. This is necessary for the exchange of information and, consequently, for drawing common conclusions based on information processed by various communities. The physical basis of information integration is nodes-hubs that provide interconnection between clusters. Without this, the work of the brain as a whole would be impossible. The integration of information is based on the relationships between clusters and is implemented by neurons from the "club of favorites".

Thus, the main essence of two phenomena evaluated in the theory of integrated information was revealed: segregation and integration of information. It can be assumed that they arise in the brain in an evolutionary way and are extremely important for the successful functioning of the brain.

Definition of the integrated information parameter Φ

Let's now consider exactly how the indicator Φ is calculated. It is necessary to derive such an indicator that will describe the degree of information generation by parts of the system and the degree of its integration. To begin with, it is necessary to accurately characterize the generation

of information. Let's imagine the system as a random variable X, which is a N-dimensional vector, where N is the number of elements (nodes) in the system. Each element of the vector shows the state of one node at a given time. Node states are considered binary — a node can be either inactive (state "0") or active (state "1"). The system evolves over time, initially being in the state X_0, then, after some time, in the state X_1, then, after the same time — in the state X_2, and so on. In this theory, time is considered discretely, at certain interva Δt [3]:

$$X_1 \underrightarrow{\Delta t} X_2 \underrightarrow{\Delta t} X_3 \ldots$$

Then the amount of information generated by the system in one step Δt can be estimated using the expression:

$$H(X_t|X_{t+\Delta t}),$$

where $H(X|Y)$ is the conditional entropy of the variable X, provided that the variable Y has been observed. Here the conditional entropy shows the amount of information needed to describe the system (complete removal of uncertainty about it) at time t, provided that the state of the system is known at time $t+\Delta t$. We can say that this indicator predicts to what extent the state of the system at time t predicts the future state of the system at time $t+\Delta t$. Now it is necessary to assess the degree of integration of information, that is, the interaction of parts of the system. To do this, Tononi and colleagues proposed splitting the system into disjoint parts, removing all connections between these parts, and calculating a simple difference between the amount of information generated by the whole system and the generated sum of its parts. This metric is called effective information (Effective Information, EI):

$$EI(X_t \to X_{t+\Delta t}, P) = \left[\sum_{i=1}^{k} H(P_{t,i}|P_{t+\Delta t,i})\right] - H(X_t|X_{t+\Delta t}).$$

Here $P = P_1, P_2, \ldots P_k$; $P_1 \cup P_2 \cup \ldots \cup P_k = X$ — splitting the system into k parts. The entropies under the sum are calculated for each part of the system, for some subset of its elements P_i. Obviously, this difference will not always be zero, the entropy of the sum of the parts is not equal to the entropy of the whole system, since probability distributions appear in the expression for conditional entropy, and the sum of distributions in general is not equal to the joint distribution. Equality is achieved only in the case of complete independence of the parts, when the separation will not change anything — then the sum of the entropies of the parts will be equal to the total entropy of the system, and the difference will be zero. This is a very important special case that helps to understand the principle of assessing the degree of integration of information — the degree of interdependence is assessed, the degree of what consequences will be from the separation of the system into parts. The greater the gain from combining parts into a system, the more information can be additionally obtained from such a combination, the more effective information will beя EI. Now it is necessary to determine which specific partitioning of the system into parts should be used for calculations. According to the definition given in [3], it is required to find a partition that will give the minimum value of EI, that is, one that will determine the most independent parts of the system. This makes it possible to determine the weak point of the system - the part of it that is minimally connected to the rest of the system and, therefore, weakly exchanges information with it. The information generated by this part is least used by the system to create general conclusions and makes the least contribution to the system. In the worst case, such a minimum value is zero, which means that it was possible to find a completely independent subgroup in the system that does not exchange information with

the rest of the parts. The found partition in IIT is called a partition with minimal information (Minimum Information Partition, MIP). And the value of the integrated information coefficient Φ is the value of the effective information EI when splitting MIP:

$$\Phi = \mathrm{EI}(X_t \to X_{t+\Delta t}, P = \mathrm{MIP}).$$

Finally, one problem needs to be fixed. If the parts into which the system is divided are very different in size, then effective information will primarily be determined by the size of the parts of the system, and not by the degree of interconnection and interdependence, as required. For example, in the case of splitting into two strongly unequal parts, the effective information will almost always be less than when divided into equal parts, since the separation of a small number of elements weakly affects the system and weakly changes the amount of information generated in it. Similarly, splitting into many parts will reduce information much more than splitting into a small number — the more parts, the more connections will be broken, the more interdependencies will disappear, and the system will generate less information. To eliminate such effects, when searching for a partition, the minimum is searched for by the normalized values of EI, where the normalization coefficient N_P is equal to the minimum possible information generated by all parts of the system:

$$N_P = (k-1) \min_{i=1\ldots k} H_{\max}(P_i),$$

where k is the number of parts in the partition, and $H_{\max}(P_i)$ is the maximum entropy of a part of the system (the maximum possible amount of information required to describe its state). For systems with binary nodes, the maximum entropy is equal to the size of the system. Thus, $N_P = (k-1)\min_{i=1\ldots k}|P_i|$ [3]. As a result, the indicator of integrated information is determined by the expression

$$\Phi = \arg\min_{P} \frac{\left[\sum_{i=1}^{k} H(P_{t,i}|P_{t+\Delta t,i})\right] - H(X_t|X_{t+\Delta t})}{(k-1)\min_{i=1\ldots k}|P_i|}, \tag{1}$$

from which it can be seen that the indicator of integrated information is measured in bits, since Φ is actually a linear combination of various entropies, which are also measured in bits. In theory, Φ non-negative, since the sum of the Shannon entropies of the parts of the system will be greater than the entropy of the whole system — more connections gives more interdependencies, which reduces the number of possible states of the system and reduces entropy. Theoretically, the maximum possible value of Φ is equal to the number of elements in the system, but in practice it is never achieved, since this requires the entropy of the whole system to be zero, which is impossible.

Practical calculation of the coefficient of integrated information

The main problem of computing by such a classical definition is finding a partition with minimal information. No assumptions and restrictions are made about the type of partitioning — theoretically, it can be arbitrary, into any number of parts and with any distribution of elements over these parts. In practice, such a partition can be found only by a complete search of all possible partitions, with the calculation of normalized effective information for each and the search for the minimum. Let's estimate the complexity of the algorithm. The number of all possible partitions of a set of N elements is determined by the Bell number B_N. It is given by

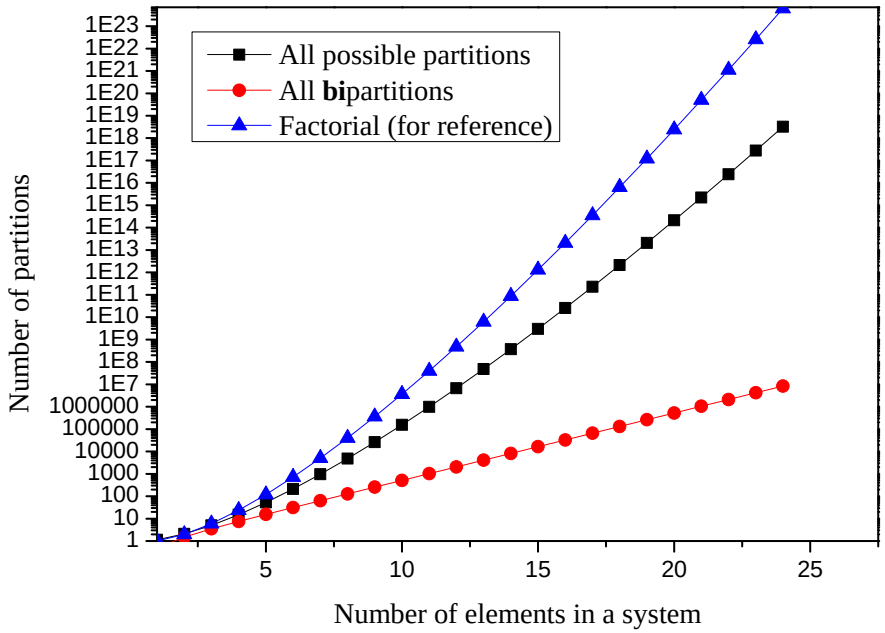

Fig. 3. A number of system partitions required to consider to find the Minimal Information Partition (MIP). Number of partitions is shown for arbitrary partitions and bipartitions; also a factorial of a number of elements is shown for reference. The y axis is in logscale. Plotted using the estimation of a number of arbitrary partitions [55] and a number of bipartitions [56]

the recurrence relation [54]

$$B_N = \sum_{k=0}^{N} C_n^k B_{k-1}.$$

Assuming for obvious reasons that $B_0 = 1$ and $B_1 = 1$, it is easy to calculate the Bell number for the following N as well. So, for example, $B_{10} = 115975$, $B_{15} = 1382958545$, a $B_{20} = 51724158235372$ (fig. 3).

The Bell number can be estimated as [55]:

$$B_N < \left(\frac{0.792n}{\ln(n+1)}\right)^n,$$

that is, it grows faster than the exponent, but slower than the factorial. Thus, even for relatively small systems, calculating Φ becomes a big problem. Modern computing power, even on supercomputers, will not be enough to determine the Φ index for systems with $N \geqslant 10$ elements. Therefore, in practice, they are often limited to splitting the system into two parts. In this case, the number of partitions that need to be sorted is determined by the Stirling number of the second kind [56], which is already growing exponentially (see Fig. 3):

$$S(N,2) = \frac{1}{2}(2^n - 1)$$

However, even in such a simplified version, the task remains very difficult. For example, for 30 neurons, when performing calculations, you will have to iterate over $S(30,2) = 5.369 \cdot 10^8$

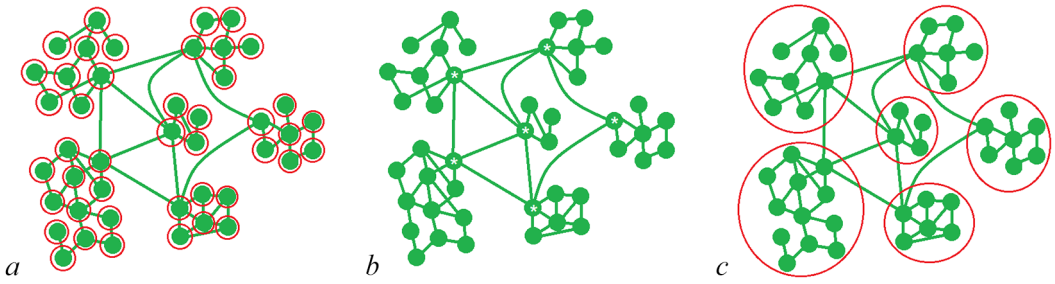

Fig. 4. Approximated partitions of a brain neural network partially replicating properties of MIP. *a* — Atomic partition, every neuron is contained in its own partition (shown by the red circles). Such partition does not correspond to an existing structure of a neural network and neglects the information processing by neuron groups. *b* — Neural network cluster structure. Neuron communities connected between each other by synapses with hubs (marked by astericks) are visible. *c* — Neural network partition by groups corresponding to neuron clusters. Changed from [21] (color online)

possible partitions. Therefore, approximate methods are needed that allow the calculation to be performed faster, obtaining Φ, not much different from the exact value. They are based on finding a partition that more or less repeats the properties of the true one (required by definition, given in IIT), but is found without a complete search. One of the first proposals was to use the so-called Atomic partition [13, 57, 58], which is a simple division of a system of N nodes into N parts, where each node is in one part (Fig. 4, *a*). This partially solves the problem, since it corresponds to the basic ideology of the theory of integrated information — calculating the gain in the operation of the whole system compared to the work of individual parts. At the same time, it is approximately assumed that such separate parts are the nodes themselves, and the presence of separate clusters (subsets) capable of generating information is ignored. This is a serious drawback of this approach, since the real structure of neural networks in the brain, as mentioned earlier, is not homogeneous, but consists of separate clusters with close connections inside (Fig. 4, *b*). Each of these clusters solves a specific task and is to some extent an independent unit, and splitting the network into separate neurons ignores this fact and does not allow us to assess the gain in information generation due to the operation of such clusters. Therefore, approximations that take into account such a cluster structure are necessary.

It can be assumed that the partitioning into such clusters is the closest to that required by the definition of Φ given in IIT (Fig. 4, *c*). Splitting into places where there are the least connections between the parts of the system (that is, splitting outside the boundaries of clusters) will lead to the least reduction in the amount of information generated by the system. Methods of statistical analysis of neural activity are used to identify such clusters. The high correlation between the series of spike events of two neurons means that with a high probability the activity of one neuron can be the cause of the activity of the other, that is, one of them has a synaptic connection with the other. The strength of the synaptic connection is determined by the magnitude of the correlation. After that, well-known methods for finding close clusters in a weighted graph are used, such as the Loven method [59], spectral clustering [60] or the weighted stochastic block model [61]. This allows you to calculate Φ in a reasonable time, getting a fairly accurate result [57].

Calculating Φ by definition (1) is very difficult in practice. Here there is a well-known problem that arises when obtaining statistical distributions — the amount of data may become severely insufficient to obtain a sufficiently accurate probability distribution. For a system of N elements, there are 2^N possible states, and it is necessary to observe each of them enough times to calculate the probability of occurrence of each state and thus obtain a probability distribution

close to the true one. In practice, problems manifest themselves in the occurrence of instability of the calculated value of Φ, when a small difference in the input data (for example, observing a slightly smaller number of states) leads to very large changes in Φ). At the same time, the normalized effective information EI/N_P for several partitions of the system practically does not differ and is near the desired minimum, and the non-normalized effective information EI for these partitions already differs significantly. A small difference in the collected states of the system at the same time leads to the fact that the desired minimum goes to another partition, which will give a very different non-normalized value of EI, and, consequently, a different value of Φ. Such instability can seriously distort the results. Therefore, in [62], a new algorithm was developed for calculations, the so-called autoregressive Φ. With its help, it is possible to calculate Φ for the data encountered in practice on the evolution of the states of each element of the system over time. The coefficient of integrated information in this case is calculated as follows:

$$\Phi_{\mathrm{AR}} = \frac{\min_M \frac{1}{2} \ln \frac{\det \Sigma(X)}{\det \Sigma(E^X)} - \sum_{k=1}^{M} \frac{1}{2} \ln \frac{\det \Sigma(M_k)}{\det \Sigma(E^{M_k})}}{L(M)}. \tag{2}$$

Here $\Sigma(X)$ — covariance matrix of the variable X, E^X — remains of a regression predicting the state of the system at time t based on the state at time $t+\Delta t$, E^{M_i} — the same residuals, but not for the whole system, but for its parts (subsamples) M_i, and $L(M) = \frac{1}{2} \ln \left[\min_k \left\{(2\pi e)^{|M_k|} \det \Sigma(M_k)\right\}\right]$ — normalization coefficient. The division into subsamples can be used any proposed in one of the approximate calculation methods. When calculating Φ in this way, problems are possible in cases when the data changes too slowly (the elements of the system are always in the state "0"or "1"), which leads to the impossibility of calculating regressions in (2). In this case, the least slightly changing element is removed from the system [14]. Since such elements encode little information, their removal does not affect the result.

There are other definitions in the literature in which some improvements are applied, such as taking into account the prediction at each stage of both the future and the previous states of the system [63], replacing the entropy difference with the distribution difference metric (Wasserstein metric) [4], and many others [64–66]. However, in practice, in the articles discussed earlier, for practical calculations on data in the form of time series (spike or EEG) they were not used because the calculation of the distribution characteristics required when using these methods is impossible in practice due to insufficient data volumes. For example, in the works [20–22], neural activity data were long time series, so the definition of «autoregressiveΦ» was used. In the work [13], the classical definition was used, since due to the small amount of data in an artificially simulated system, entropy calculation was possible. In [14], the autoregressive variant was used to calculate on a larger data set; for another data set, the classical definition was used.

Finally, to calculate Φ on neural activity data (spike or EEG), it is necessary to find the parameter Δt, which determines how many steps the amount of information generated by the system is calculated. When calculating Φ in simulated systems, where their evolution over time occurs in steps (using program code in a loop), (as, for example, in [13]), obviously the generation of information in one such step is considered. For continuous, non-discrete systems, the use of such an approach is illegal. The paper [20] provides an overview of approaches to choosing Δt for neural activity of the brain. A huge number of processes with complex time dynamics occur inside the brain at different levels, so any estimate of Δt will be approximate, and the calculation of Φ using such an estimate will not be based on the amount of information generated during the completed cycle. Processes occurring at different speeds are detected in the brain, both at the

level of individual neurons (action potentials and neurotransmitter release) and at the level of multiple neurons (total brain activity). At the moment, there are many confirmations that these processes occur periodically, and can be described either by the principle of a moving average, or autoregressive (each next state is derived from the previous one). It is necessary to estimate the time scale of such processes. It can be assumed that these periods are determined by rhythmic fluctuations of large neural groups, which are visible as rhythms on the EEG. Also, cognitive processes [67–69] seem to be periodic. According to various sources (see, review in [20]), the moment of conscious perception, for example, occurs in the period from 50 to 500 ms after the presentation of information, therefore it is useful to consider the periods of information generation by neural communities within such limits. In general, all sources agree that such processes occur on a time scale faster than a second. As a result, to determine the characteristic period for which the maximum amount of information is generated, it is necessary to iterate through all possible numbers of steps giving Δt within one second. The value of Δt, at which the maximum index of Φ is reached, and is used in further calculations. A similar approach was used in [14].

Conclusion

The basics of the theory of integrated information and its possible application in neuroscience to assess the process of acquiring new experience were considered. The methods that allow in practice to calculate the value of Φ for neural activity data are indicated. Currently, such an approach is used only to a limited extent due to the significant computational complexity for large volumes of spike data. However, existing studies show the advantages of using IT and its ability to assess the internal states of neural networks in the brain.

References

1. Tononi G. An information integration theory of consciousness. BMC Neuroscience. 2004;5:42. DOI: 10.1186/1471-2202-5-42.
2. Tononi G. Consciousness as integrated information: a provisional manifesto. The Biological Bulletin. 2008;215(3):216–242. DOI: 10.2307/25470707.
3. Balduzzi D, Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. PLoS Comput. Biol. 2008;4(6):e1000091. DOI: 10.1371/journal.pcbi.1000091.
4. Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. PLoS Comput. Biol. 2014;10(5):e1003588. DOI: 10.1371/journal.pcbi.1003588.
5. Tononi G, Koch C. Consciousness: here, there and everywhere? Phil. Trans. R. Soc. B. 2015;370(1668):20140167. DOI: 10.1098/rstb.2014.0167.
6. Massimini M, Ferrarelli F, Huber R, Esser SK, Singh H, Tononi G. Breakdown of cortical effective connectivity during sleep. Science. 2005;309(5744):2228–2232. DOI: 10.1126/science.1117256.
7. Alkire MT, Hudetz AG, Tononi G. Consciousness and anesthesia. Science. 2008;322(5903):876–880. DOI: 10.1126/science.1149213.
8. Gosseries O, Di H, Laureys S, Boly M. Measuring consciousness in severely damaged brains. Annual Review of Neuroscience. 2014;37:457–478. DOI: 10.1146/annurev-neuro-062012-170339.
9. Casali AG, Olivia Gosseries O, Rosanova M, Boly M, Sarasso S, Casali KR, Casarotto S, Bruno MA, Laureys S, Tononi G, Massimini M. A theoretically based index of consciousness independent of sensory processing and behavior. Science Translational Medicine. 2013;5(198):

198ra105. DOI: 10.1126/scitranslmed.3006294.
10. King JR, Sitt JD, Faugeras F, Rohaut B, Karoui IE, Cohen L, Naccache L, Dehaene S. Information sharing in the brain indexes consciousness in noncommunicative patients. Current Biology. 2013;23(19):1914–1919. DOI: 10.1016/j.cub.2013.07.075.
11. Searle JR. Can information theory explain consciousness? [Electronic resource] The New York Review of Books. 10 January 2013. Available from: https://www.nybooks.com/articles/2013/01/10/can-information-theory-explain-consciousness/.
12. Barrett AB, Mediano PAM. The Phi measure of integrated information is not well-defined for general physical systems. Journal of Consciousness Studies. 2019;26(1–2):11–20.
13. Edlund JA, Chaumont N, Hintze A, Koch C, Tononi G, Adami C. Integrated information increases with fitness in the evolution of animats. PLoS Comput. Biol. 2011;7(10):e1002236. DOI: 10.1371/journal.pcbi.1002236.
14. Engel D, Malone TW. Integrated information as a metric for group interaction. PLoS ONE. 2018;13(10):e0205335. DOI: 10.1371/journal.pone.0205335.
15. Albantakis L, Tononi G. The intrinsic cause-effect power of discrete dynamical systems–from elementary cellular automata to adapting animats. Entropy. 2015;17(8):5472–5502. DOI: 10.3390/e17085472.
16. Niizato T, Sakamoto K, Mototake YI, Murakami H, Tomaru T, Hoshika T, Fukushima T. Finding continuity and discontinuity in fish schools via integrated information theory. PLoS ONE. 2020;15(2):e0229573. DOI: 10.1371/journal.pone.0229573.
17. Fujii K, Kanazawa H, Kuniyoshi Y. Spike timing dependent plasticity enhances integrated information at the EEG level: A large-scale brain simulation experiment. In: 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). 19-22 August 2019, Oslo, Norway. New York: IEEE; 2019. P. 137–142. DOI: 10.1109/DEVLRN.2019.8850724.
18. Niessing J, Friedrich RW. Olfactory pattern classification by discrete neuronal network states. Nature. 2010;465(7294):47–52. DOI: 10.1038/nature08961.
19. Mazzucato L, Fontanini A, La Camera G. Dynamics of multistable states during ongoing and evoked cortical activity. Journal of Neuroscience. 2015;35(21):8214–8231. DOI: 10.1523/JNEUROSCI.4819-14.2015.
20. Isler JR, Stark RI, Grieve PG, Welch MG, Myers MM. Integrated information in the EEG of preterm infants increases with family nurture intervention, age, and conscious state. PLoS ONE. 2018;13(10):e0206237. DOI: 10.1371/journal.pone.0206237.
21. Nazhestkin I, Svarnik O. Different approximation methods for calculation of integrated information coefficient in the brain during instrumental learning. Brain Sciences. 2022;12(5):596. DOI: 10.3390/brainsci12050596.
22. Nazhestkin IA, Svarnik OE. Integrated information coefficient estimated from neuronal activity in hippocampus-amygdala complex of rats as a measure of learning success. Journal of Integrative Neuroscience. 2022;21(5):128. DOI: 10.31083/j.jin2105128.
23. Sporns O. Network attributes for segregation and integration in the human brain. Current Opinion in Neurobiology. 2013;23(2):162–171. DOI: 10.1016/j.conb.2012.11.015.
24. Brodmann K. Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues. Leipzig: Barth; 1909. 324 s. (in German).
25. Vogt C, Vogt O. Allgemeine ergebnisse unserer hirnforschung. Bd. 25. JA Barth; 1919. 190 s. (in German).
26. Sporns O, Chialvo DR, Kaiser M, Hilgetag CC. Organization, development and function of complex brain networks. Trends in Cognitive Sciences. 2004;8(9):418–425. DOI: 10.1016/j.tics.2004.07.008.

27. Hilgetag CC, Burns GAPS, O'Neill MA, Scannell JW, Young MP. Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat. Phil. Trans. R. Soc. Lond. B. 2000;355(1393):91–110. DOI: 10.1098/rstb.2000.0551.
28. Sporns O, Zwi JD. The small world of the cerebral cortex. Neuroinformatics. 2004;2(2):145–162. DOI: 10.1385/NI:2:2:145.
29. Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–512. DOI: 10.1126/science.286.5439.509.
30. Bassett DS, Bullmore ED. Small-world brain networks. The Neuroscientist. 2006;12(6):512–523. DOI: 10.1177/1073858406293182.
31. Muldoon SF, Bridgeford EW, Bassett DS. Small-world propensity and weighted brain networks. Scientific Reports. 2016;6(1):22057. DOI: 10.1038/srep22057.
32. Bassett DS, Bullmore ET. Small-world brain networks revisited. The Neuroscientist. 2017;23(5):499–516. DOI: 10.1177/1073858416667720.
33. Liao X, Vasilakos AV, He Y. Small-world human brain networks: Perspectives and challenges. Neuroscience & Biobehavioral Reviews. 2017;77:286–300. DOI: 10.1016/j.neubiorev.2017.03.018.
34. Milgram S. The small-world problem. Psychology Today. 1967;1(1):61–67.
35. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393(6684):440–442. DOI: 10.1038/30918.
36. Latora V, Marchiori M. Efficient behavior of small-world networks. Phys. Rev. Lett. 2001;87(19):198701. DOI: 10.1103/PhysRevLett.87.198701.
37. Latora V, Marchiori M. Economic small-world behavior in weighted networks. The European Physical Journal B - Condensed Matter and Complex Systems. 2003;32(2):249–263. DOI: 10.1140/epjb/e2003-00095-5.
38. Simard D, Nadeau L, Kröger H. Fastest learning in small-world neural networks. Physics Letters A. 2005;336(1):8–15. DOI: 10.1016/j.physleta.2004.12.078.
39. Lynn CW, Bassett DS. The physics of brain network structure, function and control. Nature Reviews Physics. 2019;1(5):318–332. DOI: 10.1038/s42254-019-0040-8.
40. Anokhin KV. The cognitome: Seeking the fundamental neuroscience of a theory of consciousness. Neuroscience and Behavioral Physiology. 2021;51(7):915–937. DOI: 10.1007/s11055-021-01149-4.
41. Rulkov NF, Sushchik MM, Tsimring LS, Abarbanel HDI. Generalized synchronization of chaos in directionally coupled chaotic systems. Phys. Rev. E. 1995;51(2):980–994. DOI: 10.1103/PhysRevE.51.980.
42. Aertsen AM, Gerstein GL, Habib MK, Palm G. Dynamics of neuronal firing correlation: modulation of "effective connectivity". Journal of Neurophysiology. 1989;61(5):900–917. DOI: 10.1152/jn.1989.61.5.900.
43. Boccaletti S, Kurths J, Osipov G, Valladares DL, Zhou CS. The synchronization of chaotic systems. Physics Reports. 2002;366(1–2):1–101. DOI: 10.1016/S0370-1573(02)00137-0.
44. Rosenblum M, Pikovsky A. Synchronization: From pendulum clocks to chaotic lasers and chemical oscillators. Contemporary Physics. 2003;44(5):401–416. DOI: 10.1080/0010751031000160312 9.
45. Malliaros FD, Vazirgiannis M. Clustering and community detection in directed networks: A survey. Physics Reports. 2013;533(4):95–142. DOI: 10.1016/j.physrep.2013.08.002.
46. Garcia JO, Ashourvan A, Muldoon S, Vettel JM, Bassett DS. Applications of community detection techniques to brain graphs: Algorithmic considerations and implications for neural function. Proceedings of the IEEE. 2018;106(5):846–867. DOI: 10.1109/JPROC.2017.2786710.
47. van den Heuvel MP, Hulshoff Pol HE. Exploring the brain network: A review on resting-state fMRI functional connectivity. European Neuropsychopharmacology. 2010;20(8):519–

534. DOI: 10.1016/j.euroneuro.2010.03.008.
48. Liao W, Ding J, Marinazzo D, Xu Q, Wang Z, Yuan C, Zhang Z, Lu G, Chen H. Small-world directed networks in the human brain: Multivariate Granger causality analysis of resting-state fMRI. NeuroImage. 2011;54(4):2683–2694. DOI: 10.1016/j.neuroimage.2010.11.007.
49. Stam CJ, van Straaten ECW. The organization of physiological brain networks. Clinical Neurophysiology. 2012;123(6):1067–1087. DOI: 10.1016/j.clinph.2012.01.011.
50. Kahnt T, Chang LJ, Park SQ, Heinzle J, Haynes JD. Connectivity-based parcellation of the human orbitofrontal cortex. Journal of Neuroscience. 2012;32(18):6240–6250. DOI: 10.1523/JNEUROSCI.0257-12.2012.
51. Yu C, Zhou Y, Liu Y, Jiang T, Dong H, Zhang Y, Walter M. Functional segregation of the human cingulate cortex is confirmed by functional connectivity based neuroanatomical parcellation. NeuroImage. 2011;54(4):2571–2581. DOI: 10.1016/j.neuroimage.2010.11.018.
52. Kashtan N, Alon U. Spontaneous evolution of modularity and network motifs. Proc. Natl. Acad. Sci. U.S.A. 2005;102(39):13773–13778. DOI: 10.1073/pnas.0503610102.
53. Lipson H, Pollack JB, Suh NP. On the origin of modular variation. Evolution. 2002;56(8):1549–1556. DOI: 10.1111/j.0014-3820.2002.tb01466.x.
54. Rota GC. The number of partitions of a set. The American Mathematical Monthly. 1964;71(5):498–504. DOI: 10.1080/00029890.1964.11992270.
55. Berend D, Tassa T. Improved bounds on Bell numbers and on moments of sums of random variables. Probability and Mathematical Statistics. 2010;30(2):185–205.
56. Graham RL, Knuth DE, Patashnik O. Concrete Mathematics: A Foundation for Computer Science. Reading, MA, USA: Addison–Wesley; 1994. 657 p.
57. Toker D, Sommer FT. Information integration in large brain networks. PLoS Comput. Biol. 2019;15(2):e1006807. DOI: 10.1371/journal.pcbi.1006807.
58. Mediano PAM, Seth AK, Barrett AB. Measuring integrated information: Comparison of candidate measures in theory and simulation. Entropy. 2019;21(1):17. DOI: 10.3390/e21010017.
59. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008(10):P10008. DOI: 10.1088/1742-5468/2008/10/P10008.
60. Ng A, Jordan M, Weiss Y. On Spectral Clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems 14. NIPS; 2001.
61. Faskowitz J, Yan X, Zuo XN, Sporns O. Weighted stochastic block models of the human connectome across the life span. Scientific Reports. 2018;8(1):12997. DOI: 10.1038/s41598-018-31202-1.
62. Barrett AB, Seth AK. Practical measures of integrated information for time-series data. PLoS Comput. Biol. 2011;7(1):e1001052. DOI: 10.1371/journal.pcbi.1001052.
63. Tononi G. Integrated information theory of consciousness: an updated account. Archives Italiennes de Biologie. 2012;150(2–3):56–90. DOI: 10.4449/aib.v149i5.1388.
64. Griffith V. A principled infotheoretic ϕ-like measure. arXiv:1401.0978. arXiv Preprint; 2014. 18 p. DOI: 10.48550/arXiv.1401.0978.
65. Oizumi M, Tsuchiya N, Amari S. Unified framework for information integration based on information geometry. Proc. Natl. Acad. Sci. U.S.A. 2016;113(51):14817–14822. DOI: 10.1073/pnas.1603583113.
66. Oizumi M, Amari S, Yanagawa T, Fujii N, Tsuchiya N. Measuring integrated information from the decoding perspective. PLoS Comput. Biol. 2016;12(1):e1004654. DOI: 10.1371/journal.pcbi.1004654.
67. VanRullen R. Perceptual cycles. Trends in Cognitive Sciences. 2016;20(10):723–735. DOI: 10.1016/j.tics.2016.07.006.

68. Fiebelkorn IC, Pinsk MA, Kastner S. A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. Neuron. 2018;99(4):842–853. DOI: 10.1016/j.neuron.2018.07.038.
69. Helfrich RF, Fiebelkorn IC, Szczepanski SM, Lin JJ, Parvizi J, Knight RT, Kastner S. Neural mechanisms of sustained attention are rhythmic. Neuron. 2018;99(4):854–865. DOI: 10.1016/j.neuron.2018.07.032.