Archives Italiennes de Biologie, 148: XX-XX, 2010.

Information integration: its relevance to brain function and consciousness

G. TONONI

Department of Psychiatry, University of Wisconsin, Madison, Wisconsin, USA

ABSTRACT

A proper understanding of cognitive functions cannot be achieved without an understanding of consciousness, both at the empirical and at the theoretical level. This paper argues that consciousness has to do with a system's capacity for information integration. In this approach, every causal mechanism capable of choosing among alternatives generates information, and information is integrated to the extent that it is generated by a system above and beyond its parts. The set of integrated informational relationships generated by a complex of mechanisms – its quale – specify both the quantity and the quality of experience. As argued below, depending on the causal structure of a system, information integration can reach a maximum value at a particular spatial and temporal grain size. It is also argued that changes in information integration reflect a system's ability to match the causal structure of the world, both on the input and the output side. After a brief review suggesting that this approach is consistent with several experimental and clinical observations, the paper concludes with some prospective remarks about the relevance of understanding information integration for analyzing cognitive function, both normal and pathological.

Key words

Qualia • Matching • Degeneracy • Causality • Emergence • Backconnections • Complexity

Introduction

This special issue is devoted to recent efforts at understanding the neural substrates of cognitive functions, their pathological changes, and their potential recovery. There is growing realization that such an understanding must face the challenge and the opportunity offered by the complexity of the brain: the brain is not constituted of a collection of neatly separated modules, which one could repair one-by-one just as one can repair a well-engineered car. Rather, the brain is a very large network of neurons that, while functionally specialized, have copious opportunities to interact, rather like living beings in a jungle or a society. In fact, if one allows for a few synaptic steps, virtually any neuron in the corticothalamic system can potentially interact with any other neuron. This means that understanding cognition requires a good grasp of the brain's overall anatomical, functional and effective connectivity, and that the activity of neurons in brain networks must always be interpreted in view of the overall 'neural context' (McIntosh, 2000; Jirsa et al., 2010; McIntosh et al., 2010).

This final contribution emphasizes that there is no full understanding of cognition without an understanding of consciousness. In fact, consciousness is what makes human cognitive abilities especially sensitive to context and therefore powerful in facing a world having a rich causal structure. Moreover, this contribution argues that the coexistence of functional specialization and integration in brain networks – the kind of complexity that allows for the brain's cognitive prowess and robustness – is

Corresponding Author: Giulio Tononi, Department of Psychiatry, University of Wisconsin, 6001 Research Park Boulevard, Madison, WI, 53719, USA - Email: gtononi@wisc.edu

actually responsible for generating consciousness. Below, we briefly revisit the idea that consciousness is a function of a system's capacity for information integration. We define integrated information as a fundamental quantity associated with every causal mechanism capable of choosing among alternatives, revealing a duality between causation and information. We emphasize that both the quantity and the quality of experience can be accounted for in terms of the information structures or qualia (set of informational relationships) generated by a complex of mechanisms. We briefly discuss how information integration can vary with the spatial and temporal grain size of the interactions within a system, leading to a straightforward definition of emergence. Finally, we consider how a system's capacity for information integration reflects its ability to match the causal structure of the world, both on the input and the output side. The account below builds upon and extends an approach featured in an earlier article in this journal (Tononi, 2001) and in subsequent articles (Tononi, 2004; Balduzzi and Tononi, 2008; Tononi, 2008; Balduzzi and Tononi, 2009). It concludes with some prospective remarks about the relevance of understanding information integration for analyzing cognitive function, both normal and pathological.

Consciousness and information integration

Empirical evidence has proven essential both in pointing out which aspects of brain anatomy and function are important for consciousness, and in providing a large number of facts in need of a coherent explanation. For example, we know that certain brain regions, such as the corticothalamic system, are essential for consciousness, while others, such as the cerebellum, are not. Similarly, classic experiments have shown that consciousness is associated with an activated electrocorticogram that is controlled by the reticular activating system (Moruzzi and Magoun, 1949). New experimental tools are refining our appreciation of when and where in the brain changes in neural activity are correlated with changes in conscious experience (Koch, 2004). However, there is also a need for a theoretical analysis that is both internally coherent and can account for the empirical evidence in a self-consistent, parsimonious manner. Above all, a theoretical analysis is necessary to understand what determines both the quantity and quality of consciousness generated by a neuronal network.

In previous work, it was suggested that, in a general sense, the quantity of consciousness is determined by the amount of integrated information generated by a complex of elements (Tononi, 2001, 2004). The quality of consciousness, in turn, is determined by the set of informational relationships or quale generated by the submechanisms of that system (Tononi, 2004, 2008; Balduzzi and Tononi, 2009). Briefly, integrated information is high if a system's mechanisms can generate a large amount of information, *and* this information is integrated. High *information* means that a system's causal mechanisms can specify precisely which out of a large repertoire of potential states could have caused its current state. High *integration* means that the information generated by the system as a whole is much higher than the information generated by its parts taken independently. In other words, integrated information reflects how much information a system's mechanisms generate above and beyond its parts.

A bit of theory

The idea that the essence of consciousness has to do with information integration is based on thought experiments rooted in phenomenology, which have been recounted before (Tononi, 2004, 2008). However, the notion of information integration and its corollaries are relevant for understanding complex systems whether or not one accepts their relevance to consciousness. In what follows, we briefly review how integrated information can be calculated for simple systems. This requires the concept of effective information and that of minimum information partition, which can be used to identify complexes of integrated elements. Then we describe how one can characterize the set of informational relationships generated by a complex – its quale or information structure. We also discuss briefly how integrated information varies at different levels of organization, by considering a system at the level of micro- or macro-elements. Finally, we introduce the notion of information matching as the change in

integrated information when a system is exposed to a structured environment.

Information

Consider an isolated system X composed of n units V (vertices or nodes) and k connections K (edges or links between source and target elements). The ssytem's connectivity is assumed to be a directed graph that may or may not be strongly connected (there is a path from any node to any other node). The system goes through discrete states over time steps. Units are assumed to be binary, and states are indicated as [1] (TRUE or ON) and [0] (FALSE or OFF). An *elementary mechanism* is a unit V that receives K ≤ 2 input connections from other units, performs an operation on those inputs, and outputs its new state[1]. Input/output functions defining each mechanism (transition probability matrices, truth tables) can be deterministic or probabilistic.

In what follows, we assume that any physical system in a certain state and endowed with a certain causal mechanism *intrinsically* and necessarily 'generates' information: purely by virtue of being in that state and having that mechanism, the system reduces uncertainty about which of its possible states might have caused its present state, and which might not. This information captures the "differences that make a difference" to the system itself (Bateson, 1972) and is an intrinsic, observer independent property. Contrasting with this *intrinsic perspective* is the *extrinsic perspective* of an observer external to the system: the observer can ask how extrinsic information (average surprise) is encoded, communicated or stored given the system's state and the observer's expectations (prior distribution based on observing the system for a long time), without regard for its causal mechanisms.

To evaluate the information intrinsically generated by a system, one can perturb it to reveal all its causal mechanisms (Fig. 1), by imposing all possible initial states with equal probability at time t_0. This is equivalent to treating the system as if disconnected into independent noise sources (atomic partition of the system, AP). For instance, for a system with two sensory elements and one AND gate (indicated by the symbol ∧, Fig. 1A), there are $2^3 = 8$ possible states in the *potential repertoire*, [0,0,0], [1,0,0], [0,1,0], [0,0,1], [1,1,0], [1,0,1], [0,1,1], [1,1,1], all equally likely (p = 1/8, Fig. 1B). This is the maximum entropy distribution on the states of the units of the system, also called its *potential repertoire*. Given the mechanisms and connections of the system as a whole, and the current state of its units, one can calculate, using Bayes' rule, the *actual repertoire* given the system's state $X_{i,t1}$ (Balduzzi and Tononi, 2008). This establishes the probability of previous system states (at t_0) that could have caused the actual, *current* state of the system (at t_1). For instance, if the current state of the AND gate system is [0,0,1], it could have come exclusively from prior states [1,1,1] or [1,1,0], each with p = 1/2 (Fig. 1B). The difference made by the mechanisms/connections of the system and its current state – the difference between the potential and actual repertoires – gives the *effective information* generated by the system. In formulas:

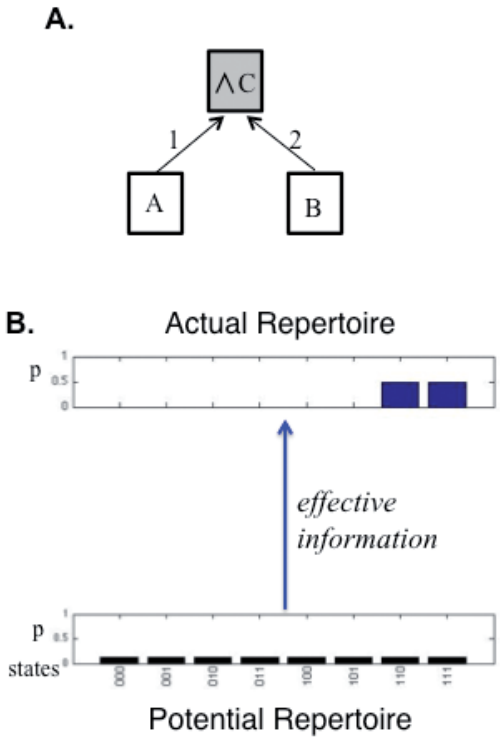

Fig. 1. - Effective information. A. System with two sensory elements and one AND gate. B. Potential and Actual repertoires.

$$ei(X_{i,t1}) = H[X_{i,t1} \,||\, X_{i,t0}/AP]$$

where AP indicates the atomic partitions of X (all units are injected with independent noise sources) and H indicates the *relative entropy* between the actual and the potential repertoires ("relative to"

is indicated by ||). Relative entropy, also known as Kullback-Leibler divergence, is a difference between probability distributions (Cover and Thomas, 2006): if the distributions are identical, relative entropy is zero; the more different they are, the higher the relative entropy. Figuratively, the system's mechanism and state generate information by sharpening the uniform distribution into a less uniform one – this is how much uncertainty is reduced. Clearly, the amount of effective information generated by a system is high if it has a large potential repertoire and a small actual repertoire, since a large number of initial states are ruled out. By contrast, the information generated is little if the system's repertoire is small, or if many states could lead to the current outcome, since few states are ruled out. For instance, if noise dominates (any previous state could have led to the current one), no alternatives are ruled out, and no information is generated.

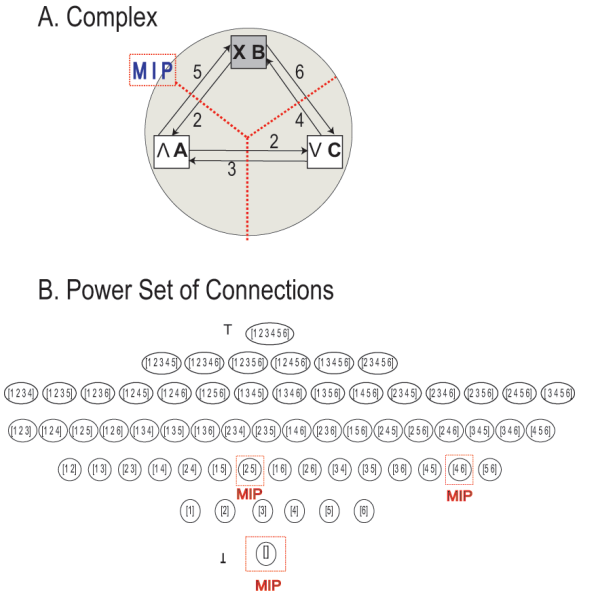

Fig. 2. - A complex, its power-set of connections, and its minimum information partitions. A. Complex. B. Power-set of connections.

Integration and complexes

Next, one must determine how much of the information generated by a system is integrated information, that is, how much information is generated by the system as a single entity, as opposed to a collection of parts. To do so, consider the set of all subsets of connections K of a system (its power-set), such as the system in Fig. 2A[2]. In the power-set diagram (Fig. 2B), the empty set (no connections) is at the bottom, all combinations containing one connection in the first layer, two connections in the second layer, and so on, ending with the set of all connections (connected system) at the top. In total, there are 2^K points (subsets) in the diagram. The points in the power-set diagram can be thought as representing the *partitions* P of the set of connections K under the total order criterion 'activation': each point partitions the K connections into an 'activated' and an 'inactivated' subset. The connections in the activated subset (TRUE or ON), which are listed in the diagram, exert causal effects. The inactivated connections (FALSE or OFF) are not listed, but they represent the complement in K of the activated ones (in Fig. 2B, subset [1] ON (second layer, left) implies that subset [2 3 4 5 6] is OFF, whereas subset [2 3 4 5 6] ON (fifth layer, right) implies that subset [1] is OFF). Inactivated connections can be thought of as "injected" with noise, so they cannot exert any causal/informational effects (this can be done by appropriately adjusting the transition probability matrix of the system).

Importantly, some of the partitions between activated and inactivated connections (*disconnecting partitions* P, or edge-cuts) yield *spatially disconnected components*. That is, inactivating those connections renders the parts of the system (subgraphs) causally and informationally independent. One can then evaluate how much information is generated by the system as a whole (the connected system) above and beyond its parts (the system disconnected into spatially independent parts along the partition P). This is given again by the relative entropy between the actual repertoire of the system X as a whole, and the repertoire generated by its parts taken independently along the partition P (as can be done by injecting noise along the edge-cut)[3]:

$$\varphi(X_{i,\,t1}) = H\,[X_{i,\,t1} \,\|\, X_{i,\,t1} \,/\, MIP]$$

Here, the symbol φ ('small' phi) stands for the amount of integrated information generated across the *minimum information partition* (MIP). This is the partition of the system into parts such that the parts, taken independently, generate the most information, therefore leaving the least information unaccounted (dotted line in Fig. 2A and B). This measures precisely the informational contribution of the

system as a whole above and beyond its parts. The average value of φ, indicated as <φ>, measures how much the information capacity of the complex as a channel (between its past and its present) is higher than that of its parts taken independently[4].

Note that there are many disconnecting partitions of a system with V units and K connections, so evaluating the MIP exhaustively is possible only for very small systems. However, knowing the underlying graph helps, since one needs to worry about disconnecting partitions, rather than about all possible partitions, and one can employ algorithms to evaluate graph components[5]. Note also that, to fairly compare different disconnections to find the MIP, it is necessary to normalize the values of effective information by the maximum possible value of φ across that disconnection. As a normalization factor, one can use the maximum number of bits (units) that can be specified through causal interactions across each disconnection, given the number of input connections available to each unit (this corresponds to the capacity of the channel across the disconnecting partition, taking the parts as sources and targets)[6].

Note also that one should consider disconnections both in space and in time. Taking again the underlying graph as a starting point, it is clear that a subset of connections can exert joint effects (and thus generate integrated information) only if there is time for perturbations to percolate through the entire subset and exert joint effects somewhere in the system. To take time into account, one can enforce causal independence between the output transmitted by a unit and the computations it performs on its inputs. In other words, one can inject noise between the computations on the inputs on one side, and the output of a unit on the other side (conditioning the output over the inputs for a given time interval). If no effective information is lost when the system is temporally disconnected for a given time interval, then the temporal disconnection is the MIP, and the system is not integrated over that time interval[7]. Again, knowing the underlying graph helps, since a subset of connections can be integrated only if the connections form paths of equal or shorter length than the number of time steps allotted.

Finally, one should measure φ values for both temporal and spatial disconnections for different subsets of connections (connected subgraphs) to find *complexes*, i.e. informational 'wholes' that are more than the sum of their parts. A *proper* complex can be defined as a set of units S with $\varphi > 0$ whose subsets R have lower or at most equal φ and whose supersets T have strictly lower φ ($R \leq S > T$, for all $R \subset S$ and all $T \supset S$)[8]. For example, consider a causal graph composed of two large parts, heavily interconnected within, and a small bridge between them (Fig. 3). After an appropriate interval, the value of φ for the MIP of each of the large parts is high, but that for the whole graph is very low. Inside each large part, smaller subsets have lower φ than the entire part. Thus, there are two complexes, the two large subgraphs. These, and only these, can be considered

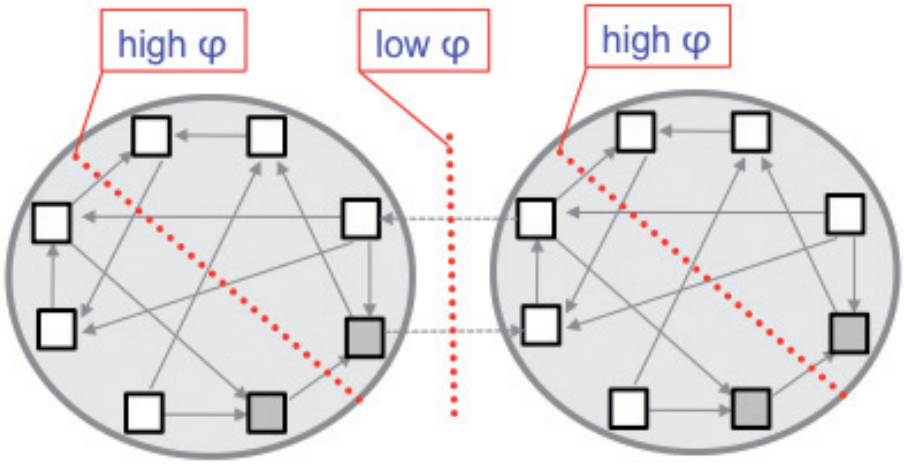

Fig. 3. - Identifying complexes based on their minimum information partition.

as individual entities: they cannot be decomposed into more integrated parts and are not part of a more integrated system. In other words, a complex is a set of elements among which integrated information reaches a maximum (as compared to any subset or superset)[9].

Irrespective of the computational problems inherent in finding the MIP for a set of connections, it is apparent that, for φ to be high, the connected system should generate much more information than any decomposition into subsets of disconnected parts (properly normalized, and at a particular temporal interval). In other words, integrated information for the MIP captures the information generated by causal interactions in the whole, over and above the information generated by causal interactions within the parts.

The information structure or quale generated by a complex

Within a complex – a whole that exists informationally above and beyond its parts – its elementary mechanisms and connections combine to generate information in specific ways: the *quale* Q or *effective information matrix* (Tononi, 2004). To characterize it, it is useful to define qualia space (Q-space) as having as many dimensions as the potential repertoire of the complex (for binary units, 2^n), and a range giving the probability of that state, from p = 0 to p = 1 (Balduzzi and Tononi, 2009). Consider again the set of all subsets of connections K of the system in Fig. 2A. Each combination of activated connections specifies an actual repertoire – the probability distribution of the states of the system – made more or less likely by those connections. The probability distributions generated by various subsets of connections are shown on Fig. 4B. In Fig. 4A, each distribution is plotted as a point in Q-space (corresponding to the tip of each arrow). In the figure, the points generated by each subset of connections are plotted in two dimensions at locations that correspond to their position in the power-set diagram (Fig. 2B). Q-space would actually have 64 dimensions (2^6), where each dimension gives the probability, from 0 to 1, of one of the 64 possible states of the system of 3 units. The empty set (no connections activated, bottom of the power-set) corresponds to the atomic partition (AP, all units are causally independent). The actual repertoire generated in this case is the same as the potential repertoire, meaning that all possible prior states are equally likely – the maximum entropy distribution. This corresponds to a point in Q with p = 1/n on all axes, called the 'bottom' ⊥ or origin of Q. On the other hand, a subset of activated connections specifies that some prior states are more likely than others, corresponding to a point in Q at p > 1/n on some axes and p < 1/n on others[10]. The actual repertoire generated by the complex with all connections activated is the top T of Q.

The "distance" or divergence (Cover and Thomas, 2006) between this point and the maximum entropy distribution – the relative entropy between the two distributions – is the effective information generated by a subset of activated connections[11], and it is represented by a *q-arrow* joining two points (repertoires) in Q-space. In Fig. 4A, the thickness of the q-arrow represents the quantity of effective information generated by the added connection and its direction in Q represents its quality. Note that activating a connection generates q-arrows that may differ in different *contexts*, namely depending on which other connections are activated. Typically, a connection will specify repertoires differently in the 'null' context, where no other connection is activated, than in the 'full' context, where all other connections are activated. In fact, the informational role of a particular connection in a complex is best seen as an entire *q-fold* in Q: the set of q-arrows that connection specifies in all possible contexts.

Just as one can define φ as the amount of integrated information generated by a *set* of connections, one can also define φ for a *subset* of connections in Q. This is the amount of information generated by that subset of connections above and beyond its parts taken independently along its minimum information partition. As before, this can be evaluated by disconnecting the parts/subgraphs causally and informationally by "injecting" noise. If φ > 0 for a subset of connections included in a complex, that subset forms a *proper* '*submechanism'* of the complex[12]. As we have seen, a set of connections that do not generate more information together than independently do not form a distinct complex. Similarly, it may be argued that subsets of connections within a complex that generate no more information together than independently do not form a distinct point in the quale[13]. This means, for example, that connections acting in distant parts of a grid (or of a topo-

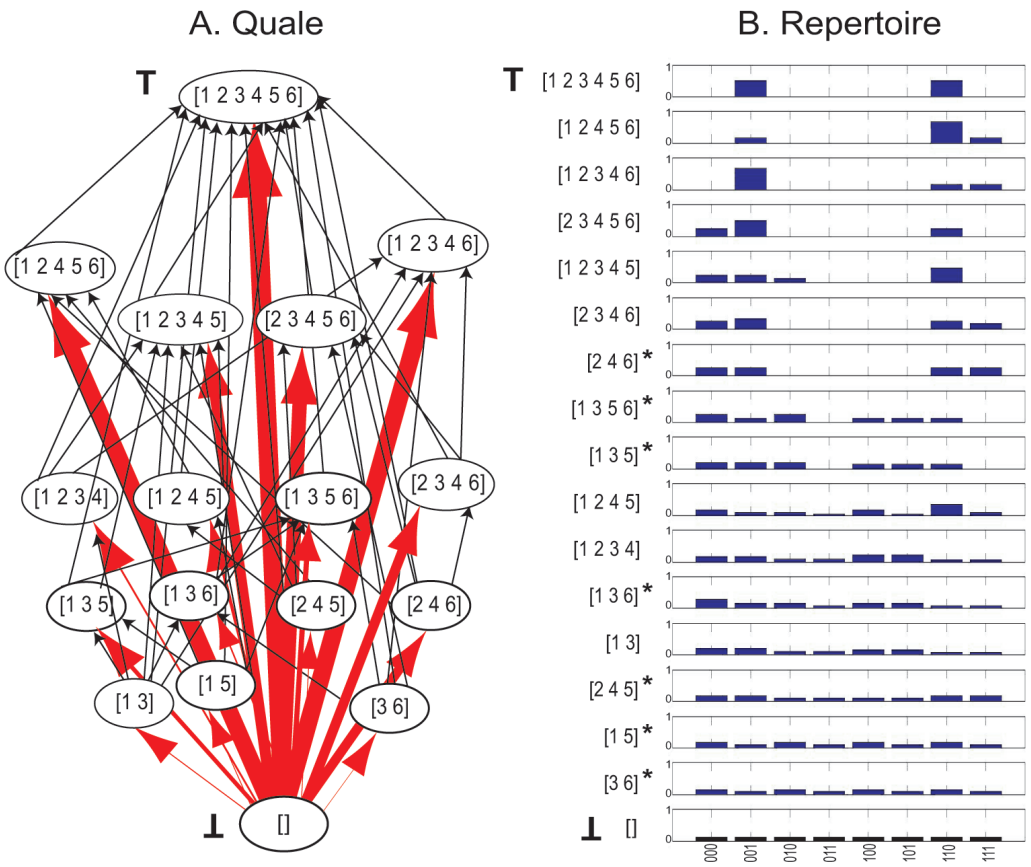

Fig. 4. - A quale, its points, and its q-arrows.

graphically organized cortical area) do not contribute to specifying an experience: they can only do so when they become integrated (capable of interacting synergistically) within a larger set of connections. The underlying principle could be stated as "no phenomenological difference without a causal difference", i.e. "no q-arrow without a mechanism"[14]. Moreover, just as one needs to attend not only to spatial but also to temporal disconnections when identifying complexes, one needs to attend to time also when evaluating informational relationships within the quale. Based on the above principle, an informational relationship in Q can be specified by a subset of connections only if there is enough time for causal influences to percolate through the subset. To take time into account, one can again enforce causal independence between the output transmitted by a unit and the computations it performs on its inputs. In other words, one can inject noise between the inputs and output of a unit (conditioning the output over the inputs for a given time interval). If for a given time interval a temporal disconnection does not change the value of effective information generated by that subset of connections, then φ for that subset at that time interval is zero, and no informational relationship is specified. On the other hand, if the time is sufficient for perturbations to percolate over paths containing all the connections in the subset, the connections can "make a difference" jointly and thus specify new points / informational relationships in the quale. Thus, by evaluating effective information values over both temporal and spatial disconnections, one can establish how the quale unfolds over time both in quantity and quality[15]. In general, one can expect that, for many systems, there will be a time interval (t_{max}) at which small phi reaches a maximum, defining the main complex. The complex will be associated with a quale containing all the informational relationships specified at t_{max} (Fig. 4)[16].

Altogether, the set of points in the quale Q, and the q-arrows linking them, are assumed to specify completely the informational quality of the experience the complex is generating (Tononi, 2004, 2008). One can think of the 'quality' of the experience, then, as the 'shape' of the quale in Q-space formed by all q-arrows together.

Once the points in the quale and the corresponding q-arrows have been defined, one can sum the values of effective information for all q-arrows in Q, which yields the *total amount of integrated information within the complex/quale*:

$$\Phi(X_{i,\,t1}) = \Sigma ei({}^{m}X_{i,\,t1,\ {}^{m\cup r}}X_{i,\,t1})$$

$$\text{for r, } m\cup r \in K \text{ and } \varphi\,({}^{m}X_{i,\,t1} \parallel {}^{m\cup r}X_{i,\,t1}) > 0$$
$$\text{over space and time}$$

where $\Sigma ei({}^{m}X_{i,\,t1,\ {}^{m\cup r}}X_{i,\,t1}) = \Sigma H[{}^{m}X_{i,\,t1} \parallel {}^{m\cup r}X_{i,\,t1}]$. This quantity, which can be called 'big' PHI, can be thought of as indicating the overall 'quantity' of consciousness in a particular experience. The value of 'small' phi for the complex, that is, the q-arrow that goes from the point corresponding to the MIP to the top of Q, represents the extent to which the full set of connections (the complex as a whole) is more than the sum of its parts (the product of their probabilities/repertoires). Small phi is thus an index of information integration for the complex as a whole, and is essential in identifying complexes. Big PHI can be considered as an index of the overall quantity of consciousness: it includes all informational relationships, not just those across the MIP; it is not subject to normalization; and it has a straightforward interpretation when considering average changes due to environmental inputs ($<\Phi(X_{World})> - <\Phi(X_{Noise})>$, see section on matching below)[17]. To the extent that the set of informational relationships can be thought of as a flow network for information within a complex, one can also think of big PHI as a measure of the overall information capacity of the network, where the cumulative capacity of all informational channels within a complex are considered. The number of bits corresponding to average big PHI ($<\Phi>$) can also be considered as a measure of the memory, in bits, that is stored in the system in terms of connections mediating informational relationships. Φ in response to a particular stimulus can then be considered as a measure of the information, in bits, that is provided by activating the system's memory. In general, this pre-existing information vastly outnumbers the (limited) information provided by the stimulus (Tononi and Edelman, 1997).

The spatio-temporal grain of information integration

The amount of information integration generated by a system, as well as the kind of informational relationship generated by its submechanisms, depend critically on the chosen spatial and temporal scale. Consider the brain: what are the elements over which one should consider perturbations and the repertoire of possible states? A natural choice would be neurons, but other choices, such as neuronal groups at a coarser scale, or synapses at a finer scale, might also be considered (not to mention molecules, atoms and so on). Yet, if integrated information is assumed to reflect consciousness, the spatial scale at which it is generated cannot be arbitrary, but must reflect somehow the actual physical mechanisms of the system at hand. This point is especially clear with respect to time. Again, integrated information can be measured in principle at many temporal scales, from picoseconds to days or longer. Clearly, however, consciousness appears to flow in a particular time range, from tens of milliseconds (Bachmann, 2000) to 2-3 seconds (Pöppel and Artin, 1988), reaching perhaps maximum vividness and distinctness at a few hundred milliseconds. Can one establish the spatial and temporal scale at which consciousness is generated in a principled manner, starting with the underlying physical mechanisms? If consciousness is indeed integrated information, then its grain should reflect the spatial and temporal scale at which integrated information reaches a maximum, given the underlying physical mechanisms (Tononi, 2004). As briefly exemplified below, integrated information can indeed behave differently for the same system at different spatio-temporal scales. Importantly, in certain circumstances a coarser scale may actually produce higher values of integrated information than a finer scale.

A relevant case is one in which an obvious distinction can be made between a macro- and a micro-level: say two interacting neurons that are coupled in a simple oscillatory circuit, vs. a set of intrinsic conductances inside each neuron that are only weakly coupled and determine its excitability stochastically. Here, the

many intrinsic channels of both neurons configure the micro-level. The two neurons configure the macro-level, where each element has just one output, represented by the axon that can spike or not (in this case, each neuron also has just one input from the other neuron). Translating this example in very simple terms, consider a system where the macro-level consists of two macro-elements (neurons {A,B}), each of which contains two micro-elements or intrinsic conductances: (A:{a1,a2}; B:{b1,b2}) (Fig. 5A). The two micro-elements within a macro-element interact weakly, for instance by being ON with a probability slightly higher than chance when the other element was ON the previous time step (slight cooperativity of intrinsic conductances in depolarizing the neuron). This is the micro-level mechanism. Let us now introduce a bridging mechanism that links the micro- and macro-level: a macro-element is ON if both its micro-elements are ON, otherwise it is OFF (if all conductances inside a neuron are ON together, the neuron spikes). The macro-level mechanism is as follows: if A was ON (neuron A spiked), B turns ON, and vice versa. Finally, the bridging mechanism that links the macro- and micro-level is as follows: a micro-element is ON, irrespective of any other inputs it may receive, if B the previous time step was ON (an incoming spike strongly depolarizes the neuron irrespective of intrinsic conductances). This set of mechanisms completely determines the behavior of the system at both the micro- and macro-level. In such a case, it can be shown that, in terms of integrated information, a complex encompassing the entire system only exists at the macro-level. Moreover, on average the value of integrated information can be higher at the macro- than at the micro-level. Thus, despite the smaller number of elements, in terms of integrated information the macro-level 'supervenes' upon the micro-level. An analogous case can be made concerning the appropriate temporal grain for information integration: depending on the actual mechanisms, a system may achieve a higher level of information integration at a coarser rather than at a finer temporal scale. For example, multiple spikes may be needed before the state of a system's units is sufficiently established to yield adequate values of effective information between subset of units[18].

Note that it is important to distinguish between levels of description and levels of interaction. An outside observer may decide to group micro-elements into macro-elements in arbitrary manners. As shown in Fig. 5B, from an extrinsic perspective one may think of grouping disconnected micro-elements {a1,a2,a3,a4} into a source macro-element A, and disconnected micro-elements {b1,b2,b3,b4} into a target macro-element B. Assume that A and B are connected by a large channel having 4 parallel lines, each copying the state of micro-element a1 onto b1, a2 onto b2, a3 onto b3, and a4 onto b4. One could then say that the effective information between macro-elements A and B is 4 bits, and it would be easy to achieve higher values by just adding micro-elements and parallel lines. However, an analysis in terms of integrated information, from the intrinsic perspective, would immediately reveal that A and B do not form complexes, and so the macro-level does not exist intrinsically. In general, a candidate macro-element can only be a grouping of micro-elements that constitutes a complex, a set of elements (here micro-elements) that cannot be decomposed into subsets of more integrated parts. Moreover, to be legitimate, each macro-element should be not only a complex of micro-elements, but one that can interact with other macro-elements through a single

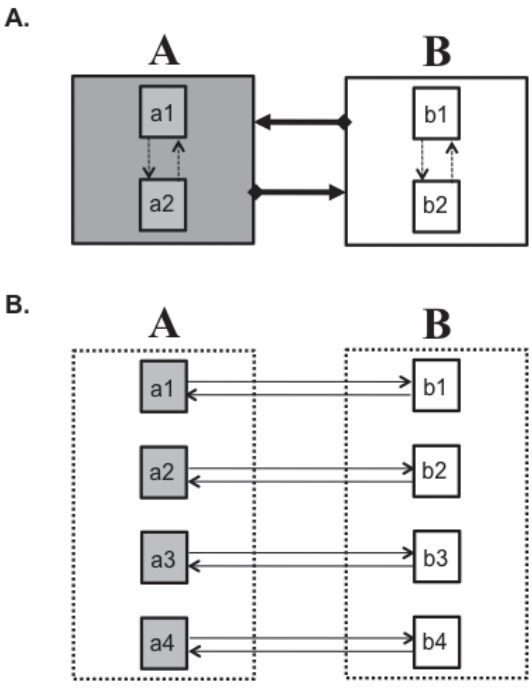

Fig. 5. - Macro- and micro-elements.

output line, just like a micro-element does. Except that, unlike a micro-element, a macro-element has an internal structure, though it is *opaque*, i.e. hidden or black-boxed behind the output line. The justification for this restrictive definition is that, if it were possible to sample the separate outputs of multiple micro-elements, it would be possible to access the micro-structure of the complex, which would ipso facto disqualify it as an 'element'.

Information matching

In any given situation, a complex of high Φ has at its disposal a large number of nested concepts – a set of informational relationships contained within a single informational structure – its quale. These nested concepts allow the complex to understand the situation in a context-dependent fashion. How can one evaluate, at least in principle, how well the integrated informational structure generated by an adapted complex, fits or 'matches' the informational structure of its environment? A measure that should be sensitive to how well an intrinsic information structure 'resonates' with the extrinsic information structure of the environment is *information matching* (M_I). This is given by the change in the average integrated information generated by a complex when it interacts with its environment, compared to when it is exposed to uncorrelated noise (a structureless environment). That is:

$$<\mathbf{M}_I> = <\Phi(X_{World})> - <\Phi(X_{Noise})>$$

In short, matching measures the average change in effective information for all the informational relationships (q-arrows) generated by a complex in a given environment. In general, one can expect that, if input matching is high, information about the environment is efficiently distributed to many subsets (parts) of a complex (cf. Tononi et al., 1996). This implies high capacity for the information channels between the input and any part of the complex – higher than expected based on the information channel between the input and the complex as a whole. It also implies that each of the parts can operate on the input in its specific way, based on its own set of submechanisms (including connections with other units mediating associations or "memories"), and efficiently provide the results of its operations to the rest of the complex. On the other hand, if a set of elements is not integrated, matching will be low, and information from the sensory input will not be distributed. This occurs, for instance, if a system is organized into parallel channels. Matching will be low also if the elements are not specialized, because all elements would perform the same operation and generate no additional informational relationships. This occurs, for instance, if a system has completely homogeneous connectivity.

In the corticothalamic system, the number of units that can be affected by an input from the environment is usually much larger than the number of input lines: rather than an information-processing device, the system is organized like a device for "interpreting" information in the light of its memories (connections). Indeed, a "snapshot" of the environment conveys little information unless it is interpreted in the context of a system whose complex causal structure, over a long history, has captured some of the causal structure of the world, i.e. long-range correlations in space and time (Tononi et al., 1996). Clearly, the connectivity of the corticothalamic context, which can be brought to bear on the interpretation of a single snapshot, is what provides this context.

Just as one can define an average measure of information integration as well as a state-dependent measure, one can define a state-dependent measure of information matching, which reflects how well a complex' casual structure resonates with a particular input. Thus, one would expect that Φ of the main complex for a monolingual English speaker would be higher when he sees an English word written in English than when he sees a Chinese word written in Chinese. A very simple example is shown in Fig. 6, where the informational structure of the quale generated by an AND gate "inflates" when the input is what the AND gate is designed to "recognize" (the coincidence "both inputs ON"). One can easily show that, on average, the value of big PHI for an AND gate will be higher when its input contains many instances of "both inputs ON" ($<\Phi(X_{World})>$, Fig. 6A), expressing a regularity in the environment that it is designed to recognize, than when such coincidences occur as expected by chance ($<\Phi(X_{Noise})>$, Fig. 6B).

Just as one can consider an input matching, one can also consider an *output matching*: for a well-adapted system, one would expect the capacity of the information channels between subsets of a complex and the

motor output to be higher when the system confronts its environment than when it is exposed to noise, and higher than expected based on the overall capacity of the channel between the complex as a whole and its outputs. Output matching is related to a previously defined measure of *degeneracy* (Tononi et al., 1999). For a given output, degeneracy is high if there are many different ways for the complex to produce that output (which ensures robustness), *and* different subsets of the complex can produce different outputs (ensuring a varied behavioral repertoire). Just like input matching implies that parts of the system have greater than expected access to the input, output matching implies that parts of the complex have higher than expected *access* to the output. It should also be pointed out that high output matching implies that the same part of a complex can produce different outputs in different contexts (pleiotropy), as is generally the case for any mechanism in a quale of high Φ. It is worth noting that total integrated information itself can be considered as a generalization of input and output matching, in that it is maximized when, for any channel between source and target elements within a complex, the capacity for partial channels (from some source elements, taken as inputs, to some target elements, taken as outputs) is higher than one would expect from the total channel capacity, implying once again that parts of the complex have high access to other parts.

Based on theoretical considerations and supported by simple simulations (Tononi et al., 1996, 1999), it is expected that both input and output matching should increase when a system adapts to an environment, meaning that spatial and temporal correlations representative of the causal structure of the environment are incorporated in the connectivity of the system by natural selection and by learning mechanisms. It is also expected, and supported by simple simulations, that an increase in matching towards an environment having a rich, integrated causal structure should lead to increased integrated information, and therefore to an increase in consciousness.

Some implications

Based on the account presented above, a physical system endowed with causal mechanisms generates an integrated information structure (complex/quale) to the extent that it cannot be decomposed into informationally more integrated parts. This notion

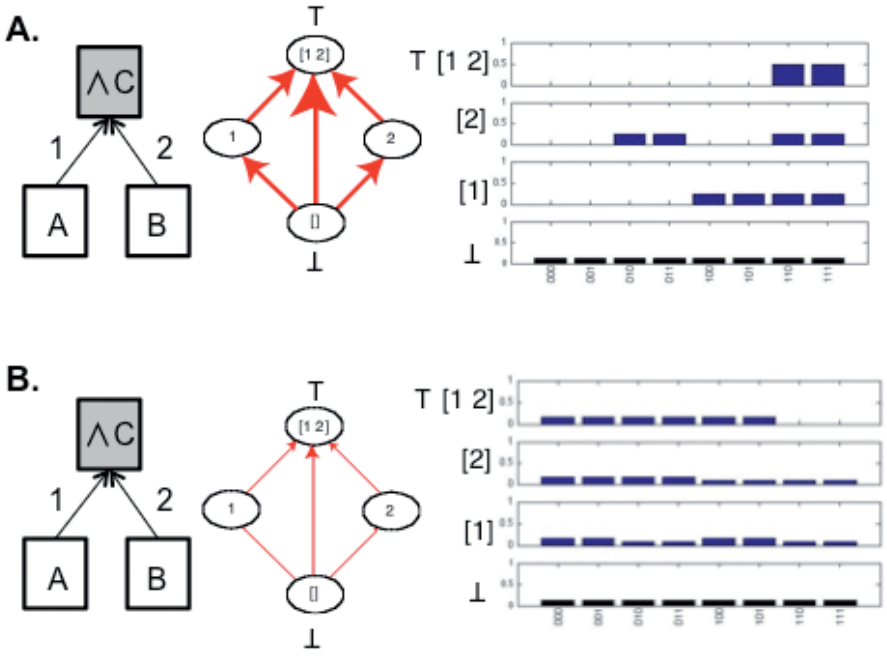

Fig. 6. - Information matching.

has several implications, both of a general nature and specifically relevant for the neuropsychology of consciousness. Below we will discuss a few of these implications, starting with the duality of causation and information.

Duality of causation and information, downward causation, and emergence

The approach taken here is that a set of connected mechanisms in a given state (*causal connectivity*) is *necessarily* or *intrinsically* associated with a certain information structure (*informational connectivity*), which it completely specifies or 'generates'. The information structure is generated if and only if the causal mechanisms are in working order, that is, they can produce different outputs as a function of their inputs, meaning that a mechanism can 'choose'. This implies that, if there is causation, there is information and, conversely, that *there is no information without causation*. It also implies that causation works *forwards* in time, in the sense that a mechanism determines the next output based on its previous input. Conversely, information works *backwards*, in the sense that the output state determined by the mechanisms generates information about the input states. If consciousness can indeed be identified with integrated information (Tononi, 2004), the *duality of causation and information* provides theoretical support for the notion that consciousness plays a causal role. In short, such a role is to be invoked whenever the causal effects of the whole (a complex) cannot be reduced to the effects of less integrated parts. In this sense, the duality between causation and information also provides support for the idea that progressively higher levels of information integration/matching, and thus of consciousness, may be an advantageous trait under natural selection. Naturally enough, an organism with a brain generating high $<\Phi>$ in an environment rich in long-range spatial and temporal correlations, can respond to environmental situations with highly context-sensitive actions. In this way it can be far more flexible than an organism equipped with a set of informationally separated processors, each of which has limited scope and understanding of the situation it finds itself in.

A useful way of contrasting systems yielding high and low information integration is to compare a response mediated by a conscious corticothalamic main complex with one mediated by a reflex arc. Say the task is to respond yes if a light is on and no if it is off (cf. the photodiode thought experiment in (Tononi, 2004)). For a reflex arc, the underlying causal matrix (transition probability matrix, truth table) includes just a small set of neurons and connections, the nodes and connections in the reflex arc. Its information 'dual'[19], the effective information matrix, would be equally small: the corresponding complex/quale would contain only a few informational relationships. For a conscious human performing the same task, instead, the relevant causal matrix would be extremely large, including a large portion of the corticothalamic system. The effective information matrix would be equally large: the corresponding complex/quale would contain a very large number of informational relationships. This complexity may be ignored when examining how the task is performed from an extrinsic perspective, say that of a neurophysiologist looking for the neurons that are activated when performing the task: one may single out a causal chain 'inscribed' on top of the corticothalamic complex and represented by the neurons that fire, from a photoreceptor in the fovea to a motoneuron, when the task is performed, and ignore the rest of the system. However, what is missed in such an extrinsic, observational approach is the large set of 'counterfactuals' in the causal matrix. In the case of the corticothalamic main complex, as opposed to the reflex arc, the silent neurons matter: if they had fired, in any of innumerable combinations, rather than having remained silent, the output would have been different. In other words, in a complex, it is just as important that some neurons fire as that the others do not, whereas in the reflex arc there are no other neurons that could affect the end result. The tendency to consider that only neurons that fire 'cause' effects, or generate information, is natural enough, but it is insufficient when dealing with an integrated system.

This perspective on the duality between causation and integrated information also has implications for the notion of '*downward causation*' (Campbell, 1974; Sperry, 1980; Szentagothai, 1984). There is no need for any kind of magical downward causation. However, there is abundant room for real downward causation. Whenever the whole is informationally more than the sum of its parts ($\Phi > 0$), then necessarily its 'dual', the causal matrix of the

complex, cannot be reduced to the sum of its causal submechanisms, and the output is necessarily a consequence of the whole. Even the previous example – the discrimination between light and dark performed by the main corticothalamic complex – turns out to involve downward causation. By applying perturbations to the corticothalamic system, it would become apparent that the "causal funnel" of a motoneuron uttering the word 'yes' involves the entire main complex: in other words, its output might have been different not only if the neurons prior to it in the causal chain that had fired had instead not fired, but also if neurons that were silent had instead fired. In this sense, then, the output – the state of the motoneuron uttering 'yes' – is determined by what the whole does (or does not do, which is the same): a case of downward causation.

Finally, this approach leads to a straightforward conceptualization of '*emergence*'. A first notion of emergence (*'weak' emergence*) is implied in the definition of integrated information as information generated by a mechanism that is not reducible to subsets of independent submechanisms. Any complex that is informationally (and therefore causally) not reducible constitutes a whole that 'emerges' above and beyond its parts. A second form of emergence (*'strong' emergence*) is implied by the idea that a complex/quale exists at the particular spatiotemporal grain at which Φ is maximal. As briefly discussed earlier, if a system generates higher integrated information at the level of macro-elements rather than at the level of micro-elements, than the macro-level of organization can be said to strongly 'emerge' upon the micro-level. This conclusion is not just a matter of semantics, but it captures precisely the fact that a complex may exist at the level of macro-elements, and not exist at all (i.e. it disintegrates) at that of micro-elements. Similar considerations can be made with respect to emergence in time and, more generally, for any kind of structure (informational relationships) that is 'manifest' in the quale at the macro-level and remains 'hidden' at the micro-level[20].

Experience as a set of informational relationships

We have argued that the quantity and quality of a particular experience are completely specified by the shape of the corresponding quale. This is the set of all informational relationships specified by the mechanisms and submechanisms (integrated subsets of connections) belonging to a complex, where a complex is a set of elements that cannot be decomposed into more integrated parts. The particular way the q-arrows are organized in Q-space, then, corresponds to the quality of consciousness, while the quantity of consciousness corresponds to the sum of the effective information values for all q-arrows (Φ)[21]. It is important to realize that, in this perspective, being conscious of even a simple stimulus, say a square in the left lower field of vision, requires a large set of informational relationships (notwithstanding the impression that the amount of information that needs to be 'encoded' is relatively small). This is because, in this perspective, the full 'understanding' of the stimulus requires the joint specification, within the same quale, of many points or probability distributions, yielding nested concepts and informational relationships. These points include probability distributions specifying that the invariant 'square' is present (a repertoire specifying particular configurations of inputs that are compatible with a square, irrespective of its position in the visual field), as well as repertoires specifying its actual details (where each edge is). Moreover, this 'vertical' compositionality must be complemented by the 'horizontal' specification of what a square is not, that is, by a large number of points specifying that alternative invariants are absent (repertoires specifying 'not a triangle', 'not a circle', 'not a face' and so on). Only if all these points (and many more) are generated within the same quale, so that the quale contains all the relevant informational relationships, can one say that a single complex 'understands' a square. In the present context, 'understanding' is one and the same thing as 'seeing' consciously.

It is also worth pointing out that all informational relationships generated by mechanisms inside a quale/complex at its privileged spatial and temporal grain size are '*manifest*', that is, they contribute to specify experience and make it what it is. By contrast, informational relationships generated by mechanisms at other spatial and temporal grains, or outside the complex, remain phenomenally '*hidden*' or unconscious. Nevertheless, mechanisms not belonging to the quale (and which are thus hidden or unconscious from its perspective) can influence its functioning. This can happen, for instance, through

units that are shared between a main complex and smaller complexes that serve as input or output channels or loops carrying out local computations, whose internal informational structure remains isolated from that of the main complex (Tononi, 2004).

Phenomenal and access consciousness

It has been suggested that the structure of experience is much richer than what can be reported (either in words or actions, (Block, 2005)). This distinction has been criticized on various grounds, chiefly because aspects of consciousness that cannot be reported would be undetectable by definition, and thus inaccessible scientifically. However, in the present perspective the distinction has merit and, at least in principle, it could be made precise. As we have seen, there is a point in the quale only if there is a mechanism for it, so all points, and thus aspects of experience, have a definite causal basis. However, points in the quale can be more or less easy to access. When a set of highly integrated connections converge onto a single unit (or onto a few similarly specialized units), they can be said to implement an '*explicit concept*', which can be thought of as a highly integrated informational relationship that is easy to access (say, the concept of a "square" or a "face"). In other words, a unit or local group of neurons implementing an explicit concept can be thought of as a locus for convergence of a large submechanism. In this case, access is comparatively easy: a question triggering a report will need to access a single convergence point that is already implemented within the complex. Indeed, the profusion of backconnections in cortical circuits (see below) suggests that it should be possible for higher areas to dynamically configure an effective path leading from the unit representing an explicit concept to units in motor cortex that will produce the appropriate answer. By contrast, if a point in the quale is specified by a combination of multiple units (i.e., it is not an explicit concept), then access is more difficult, as it would require dynamically configuring a path from multiple, distributed sources to the output units. Also, dynamically configuring access paths to a motor output may be a limited-capacity process, meaning one that can access only a few units at a time, and by the time access has been achieved, the quale has changed (for instance, in a Sperling task). This limited capacity in the report does not mean, however, that the informational relationships constituting the quale were not there. Indeed, by demonstrating 'intelligent', context-sensitive performance, it should be possible to show that informational relationships that may not be reported explicitly in isolation, have nevertheless been engaged and can make an observable difference to behavior. Also, such informational relationships would be obvious when comparing multiple scenes that would look different, though the differences may not be easy to verbalize. Not to mention that, if the informational relationships are available in the quale, over time plastic mechanisms can configure a synaptic path for direct access.

Information integration, consciousness, and neuroanatomy

As discussed elsewhere, considering the brain's capacity for information integration can account for several empirical observations concerning consciousness (Tononi, 2004; Balduzzi and Tononi, 2008). For example, by using computer simulations, it is possible to show that high integrated information requires networks that conjoin functional specialization (due to its specialized connectivity, each element has a unique functional role within the network) with functional integration (there are many pathways for interactions among the elements). This kind of architecture is characteristic of the mammalian corticothalamic system. As documented by innumerable findings, different parts of the cerebral cortex are specialized for different functions, from the level of lobes to that of areas, groups of neurons, and perhaps individual neurons. At the same time, a vast network of connections allows these parts to interact profusely. Some 'superhighways', such as the dense mesial connectivity revealed by diffusion spectral imaging (Hagmann, et al., 2008) may constitute the "backbone" of a corticothalamic main complex. So it is fitting that the corticothalamic system is precisely the part of the brain that, if severely impaired, causes a loss of consciousness. Conversely, information integration is low for systems that are made up of small, quasi-independent modules. This may be why the cerebellum, despite its large number of neurons, does not contribute much to consciousness.

As we have seen, high average integrated information means that the information capacity of the

system, considered as a channel between the past and the present (and by inference, from the present to the future), cannot be reduced to the sum of the capacities of independent channels. This implies that portions of the brain whose connectivity resembles that of parallel lines, rather than a highly convergent/divergent network complemented by lateral interactions, are ill-suited to forming large complexes of high φ. For instance, part of the dorsal cortical stream involved in action control may differ from the ventral stream in this respect (Milner and Goodale, 1995).

Certain neuropsychological observations, especially those related to disconnection syndromes, fit naturally within the framework of information integration (Tononi, 2004). In line with many studies of split brain patients (Gazzaniga, 2005), simulations show that a "callosal" cut produces, out of a large complex corresponding to the connected corticothalamic system, two separate complexes. Functional disconnections may also lead to a restriction of the neural substrate of consciousness, as is seen in neurological neglect phenomena, in psychiatric conversion and dissociative disorders, and possibly during dreaming and hypnosis. It is also likely that certain attentional phenomena may correspond to changes in the composition of the main complex underlying consciousness.

Computer simulations show that units along multiple, segregated incoming or outgoing pathways are not incorporated within the repertoire of a dominant 'corticothalamic' complex. This may be why neural activity in afferent pathways, though crucial for triggering this or that conscious experience, does not contribute directly to conscious experience; nor does activity in efferent pathways (perhaps starting with primary motor cortex), though it is crucial for reporting each different experience. Also, cortical and subcortical cycles or loops implement specialized subroutines that are capable of influencing the states of the dominant corticothalamic complex without joining it. Such informationally insulated cortico-subcortical loops could constitute the neural substrates for many unconscious processes that can affect and be affected by conscious experience (Baars, 1988; Tononi, 2004), such as those that enable object recognition, language parsing, or translating our vague intentions into the right words. A relevant question is whether parallel loops through basal ganglia implement informationally insulated subroutines. Related questions concern primary sensory cortices. Are they organized like massive afferent pathways to a main complex higher up in the cortical hierarchy (Crick and Koch, 2003), or are they involved in keeping the main complex integrated through back- and lateral connections (see below)? Similarly, how does the connectional organization of prefrontal cortex, or of the hippocampal formation, impact their contribution to the main complex?

A role for backconnections

An intriguing question concerns the role of backconnections for information integration. Backconnections are at least as numerous as forward connections but they are thought to modulate, rather than drive, the activity of their target neurons. Also, backconnections terminate profusely in supragranular layers, where NMDA synapses are abundant. Such synapses may implement coincidence detection (multiplicative interactions) between feedforward activation from lower levels and reentrant activation from higher levels (Tononi et al., 1992). Functionally, backconnections are clearly implicated in mediating attentional effects, and probably play a role in learning by making plastic changes in cortical networks context-sensitive.

These functions of backconnections may be part of a more general role in facilitating the integration of different brain areas (Tononi et al., 1992). Consider for example a unit high-up in the visual hierarchy (inferotemporal cortex, IT), which responds in a position invariant manner to, say, a particular face ("JA", top left in Fig. 7), as opposed to units responding to different invariants, such a house). Clearly, such unit has a 'neural' receptive field that encompasses units over most of primary visual cortex (V1, thick arrows converging upwards in Fig. 7), whose outputs converge on it after several synaptic steps mediated by units in intermediate visual areas (V2, V3, V4 etc.). By contrast, the receptive field of units in early areas is topographically restricted, and these units respond to local details (edges etc.). The face unit can be said to implement the general concept "JA's face", wherever it might be in the visual field. Categorizing the general concept (invariant) "JA" is useful, as it will typically call for certain sets of behaviors (e.g. approach), irrespective of

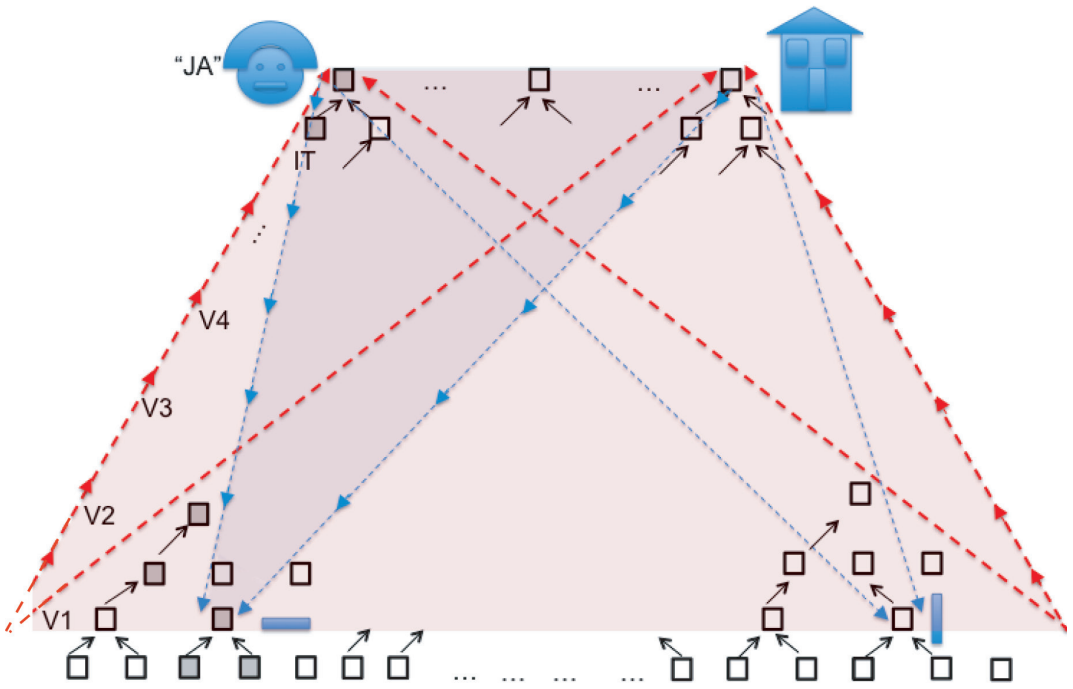

Fig. 7. - Backconnections and the integration of general and particular concepts.

location. However, the system also needs to "bind" the general with the particular: in addition to the general concept "JA", it needs the more specific concept "JA on the far left, mouth open", etc., in order to refine the appropriate behavioral output. Although units in IT implement the general concept "JA", they know little about the details such as "far left position, mouth open" (otherwise, they could not extract the invariant). On the other hand, units in early visual areas know the local details (the location and precise contour of the face and mouth), but they cannot know that it is a face and whose face it is (since these are topographic invariants).

Consider now what might happen over time in terms of complexes/qualia when JA appears in the visual field. Units in V1 begin to respond to local details (edges etc.) at time t_1. However, at t_1 these units are not properly integrated: for example, units responding to edges in the far left cannot establish causal/informational relationships with units whose receptive field is on the far right, as there are no direct links among them. At t_5, say, units in IT begin responding to JA, their specificity for JA's face and their position invariance over the entire visual field being constructed over several feedforward synaptic steps (just as important, of course, units with different specializations remain silent). The main complex will include all of the intervening units, and the quale will include informational relationships that encompass the entire visual field, such as the general concept "JA's face, wherever in the field". However, more specific concepts, such as "JA's face, on the far left, mouth open", will not yet be available. A clever way to construct such more specific concepts, rather than developing "grandmother cells" higher up, is to recruit units in lower areas with a backward sweep along backconnections (Fig. 7, thin arrows diverging downwards). Through multiplicative interactions made possible by NMDA receptors, backconnections can turn units at low levels in the visual hierarchy into coincidence detectors whose "non-classical" receptive field (Schwabe et al., 2006) can be expanded to cover potentially much of the visual field (in Fig. 7, a unit in V1 can be seen as the tip of a pyramid – the non-classical receptive field – that diverges upwards to encompass all of IT, and from there diverges downwards to encompass all of V1). From a graph-theoretical perspective, while units in lower areas may be separated by a large number of synaptic steps in the lateral

direction, the convergent/divergent arrangement of forward/backconnections establishes a short path (a "short-circuit"), going first up and then down the visual hierarchy. Causally, lower units become rapidly "reachable" in principle by a large portion of the visual system. Informationally, they can cooperate widely to specify units higher up in the visual field. Also, units that would otherwise be disconnected or separated by many synaptic steps become integrated ($\varphi > 0$), and can thus specify new points in the quale, corresponding for instance to extended contours etc. Finally, by interacting among themselves and with higher level units, lower units can now specify points corresponding to highly specific concepts such as "JA's face, on the far left, mouth open", and thus enrich the structure of the quale with new informational relationships (e.g. that the open mouth belongs to JA).

More generally, backconnections are likely to play a role, together with lateral connections, in "integrating" neural mechanisms separated by multiple synaptic steps. In this way, a large number of informational relationships can unfold within a few hundred milliseconds: just think of the rapidly growing number of neighborhood (distance) relationships that are established in a topographically organized grid as soon as interactions become possible among progressively more distant parts. The resulting informational structure in the quale would constitute nothing less than the spatial layout of visual experience, and would enable many of the classic Gestalt laws of perceptual grouping (Tononi et al., 1992; Tononi, 2008).

Although this brief sketch can only begin to address what is undoubtedly a very complex process, both causally and informationally, it is consistent with the notion that a conscious percept "unfolds" over time microgenetically, often going from the general to the particular (Bachmann, 2000). It is also consistent with evidence suggesting that a "backward sweep" through the visual system may have to occur before a conscious percept is fully formed (Cowey and Walsh, 2000; Pascual-Leone and Walsh, 2001; Silvanto et al., 2005; Lamme, 2006; Tononi and Laureys, 2009). Data from anesthesia experiments also indicate that loss of consciousness may be more closely associated with disruption of front-to-back, rather than back-to-front functional connectivity (Hudetz, 2006; Imas et al., 2006). There are also multiple observations indicating that early visual areas may contribute to experience (Cauller, 1995; Pollen, 2008), and that supragranular layers may be especially important. Finally, as was mentioned above, backconnections are ideally suited to configuring flexible paths for accessing points within the quale for conscious report.

Information integration, consciousness, and neurophysiology

Even within the same neuroanatomical structure, information integration can change drastically depending upon the mode of functioning of neurons and connections. For example, computer simulations indicate that the capacity to integrate information is reduced if neural activity is extremely high and near-synchronous, due to a dramatic decrease in the repertoire of discriminable states (Balduzzi and Tononi, 2008). This reduction in degrees of freedom could be the reason why consciousness is reduced or eliminated in absence seizures and other conditions during which neural activity is both high and synchronous (Blumenfeld and Taylor, 2003).

Perhaps the most obvious example of a marked change in consciousness depending on neurophysiological changes is the fading of consciousness that occurs during certain periods of sleep. Subjects awakened in deep NREM sleep, especially early in the night, often report that they were not aware of themselves or of anything else, though cortical and thalamic neurons remain active. Awakened at other times, mainly during REM sleep or during periods of lighter NREM sleep later in the night, they report dreams characterized by vivid images (Hobson et al., 2000). From the perspective of integrated information, a reduction of consciousness during early sleep would be consistent with the bistability of cortical circuits during deep NREM sleep. Due to changes in intrinsic and synaptic conductances triggered by neuromodulatory shifts (e.g. low acetylcholine), cortical neurons cannot sustain firing for more than a few hundred milliseconds, and invariably enter a hyperpolarized down-state. Shortly afterwards, they inevitably return to a depolarized up-state (Steriade et al., 2001). Indeed, computer simulations show that values of integrated information are low in systems with bistable dynamics (Balduzzi and Tononi, 2008). Consistent with these observations, studies using TMS in conjunction with high-density EEG

show that early NREM sleep is associated either with a breakdown of the effective connectivity among cortical areas, and thereby with a loss of integration (Massimini et al., 2005, 2007), or with a stereotypical global response suggestive of a loss of repertoire and thus of information (Massimini et al., 2007). During REM sleep, by contrast, effective connectivity recovers, just as consciousness does (Massimini et al., 2010).

A recent study has shown that a breakdown in effective connectivity, and thus of information integration, also occurs when general anesthetics produce loss of consciousness (Ferrarelli et al., 2010). Studies using multielectrode recordings in animal are also consistent with this perspective (Hudetz, 2006). A review of the mechanisms underlying unconsciousness during anesthesia concluded that much of the available evidence supports the notion that different anesthetic agents may share a common final path consisting in the disruption of information integration in a main corticothalamic complex (Alkire et al., 2008).

Finally, studies are in progress to evaluate information integration in vegetative and minimally conscious patients (Rosanova et al., unpublished). Some of these patients raise difficult questions about our ability to assess consciousness behaviorally (Owen et al., 2009). Because of sensory deficits or motor impairments, some brain injured patients recover consciousness but are unable to signal it behaviorally. Other patients still may be conscious but are unable or unwilling to follow instructions due to cognitive disturbances, pain, fatigue, or lack of motivation. Clearly, it would be helpful to develop methods to assess the level of consciousness in ways that do not depend on a subject's ability to understand or carry out instructions. Finally, what can one conclude when confronted with patients in whom an isolated cortical region shows sustained activity and even responsiveness to sensory input, in the absence of behavioral signs of consciousness (Schiff et al., 1999)? It would seem that a proper understanding of the quantity and quality of the experience that may be generated by the isolated neural mechanisms that remain functional, in the absence of the context provided by the rest of the cortex, can only be obtained with the help of a theoretical characterization of what consciousness is and how it can be generated.

Information integration and input matching

As we have seen, information matching measures the extent to which the causal/informational structure of a complex 'resonates' with the causal/informational structure of its environment. Matching can be evaluated both on the input and output side. In the first case, the change in $<\Phi>$ when a complex is exposed to its environment also reflects how well effective information from an input is distributed to different subsets of the complex. In the second case, high average effective information between different subsets of the complex and its output to the environment reflects high degeneracy: the same output can be produced by many different subsets, and at the same time different subsets can produce different outputs. These concepts are potentially relevant both in a general sense and for their neuropsychological implications.

At a general level, the present theoretical framework predicts that, to the extent that a complex matches its environment, it can do well at context-sensitive decision-making, much better than a collection of independent modules. It is also better suited at context-sensitive learning. Indeed, the ability to 'mold' a complex system as an integrated whole may be an advantage that selectional mechanisms (natural or neural) have over modular, engineered systems in 'designing' intelligent, context-sensitive systems. In essence, engineering has been extremely successful at designing systems that can perform tasks that can be decomposed (divide and conquer strategy). This strategy works if the task domain is essentially "the sum of sub-tasks". However, this strategy is showing its limits whenever a task cannot be decomposed into the sum of sub-tasks. Then the task cannot be solved by a collection of independent modules, but only by a system that works as a single entity (integrate and relate strategy). This information integration problem is related to the frame problem in artificial intelligence (difficulty of capturing the many necessary preconditions of a given action or infer its possible consequences). Similar problems occur in designing complex control systems, operating systems, robots, neural networks etc. What these hard problems have in common is that they require great context-sensitivity, because a choice that may be correct in a narrow domain often turns out to be wrong in a broader domain, which is what biological organisms, and especially brains, are extremely good at.

In a neuropsychological context, developing experimental ways of assessing matching and degeneracy could be informative both about the powers of the healthy brain, and about its potential for recovery of function after lesions. Input matching could be estimated on neuroimaging data (e.g. fMRI data) as long as one could employ an approximate measure of integrated information. One would have to compare indices of total integrated information when the brain is exposed to noise (such as TV 'snow', acoustic white noses etc.) and when the brain is exposed to the environment it is adapted to (such as movies). The comparison could be made in stages, by progressively removing statistical structure from an input stream such as a movie. One could also compare how well 'matched' a subject is to different environments or tasks. It would also be possible to exploit the difference in indices of integrated information between the waking condition and the dreaming condition. Conversely, approximate measures of matching could be used to estimate integrated information, and by extension consciousness, both across individuals and across species, since the maximum value of matching for a given brain is likely to be limited by its value of $<\Phi>$. Such an approach may be particularly useful when dealing with pathological conditions, both during development and after brain lesions.

Information integration and output matching (degeneracy)

The concept of output matching is closely related to the notion of *degeneracy*. Originally, degeneracy in a biological context was defined as the ability of elements that are structurally different to perform a similar function (Edelman and Mountcastle, 1978). Degeneracy should be distinguished from redundancy, which occurs when the same function is performed by identical elements. The key difference is that, if elements are structurally different, they may produce similar outputs in certain contexts, and different outputs in different contexts. The notion of degeneracy was developed and extended in an information theoretical context (Tononi et al., 1999). Briefly, degeneracy was defined as the amount of effective information (i.e. causal information), with respect to a set of outputs, which is shared among all subsets of a system. As shown by computer simulation, degeneracy is high for systems in which many different elements can affect the output in a similar way and at the same time can have independent effects. By contrast, degeneracy is low both for systems in which each element affects the output independently and for redundant systems in which many elements can affect the output in a similar way but do not have independent effects. As such, degeneracy is likely a key property allowing the corticothalamic system to adapt to structural damage.

To make the notion of degeneracy useful in the context of neuropsychiatry and recovery of function, once can introduce the concept of *degeneracy maps*, which can in principle be determined using available neuroanatomical and neurophysiological approaches. As we have seen, because of the convergent-divergent connectivity of the brain, large numbers of neuronal groups are able to affect the output of any chosen subset of neurons in a similar way. For example, a large number of different brain structures can influence, in series or in parallel, the same motor outputs, and after localized brain lesions alternative pathways capable of generating functionally equivalent behaviors frequently emerge spontaneously or due to plastic changes superimposed on preexistent anatomical pathways. It would thus be important to establish, prior to the occurrence of a brain lesion, the degree to which the output of a given brain area can be affected by other brain areas, that is, to obtain a *degeneracy map* for that area's output. The underlying rationale is that, to the extent that there is degeneracy with respect to that area's output, there is room for recovery of function, either immediately or through the strengthening of available connections through plasticity (Perfetti et al., 2010; Price et al., 2010; Sarasso et al., 2010; Yourganov et al., 2010).

Several kinds of degeneracy maps can be distinguished. For example, one can define a degeneracy map based on behavior. This is the set of brain areas that can influence a behaviorally defined function. Under baseline circumstances, a primary set of areas is responsible for the production of a given function. Other areas, while potentially capable of contributing to that function, need not be involved. However, in certain contexts, such as increasing task difficulty, or after brain damage, these other areas may be recruited to support that function. Degeneracy maps based on effective connectivity are the set of brain areas that can independently influence the outputs

of a target brain area. Here, the emphasis is not on a specific behavior, but on neural activity within a brain area known to be important for that behavior and related ones. For example, the target area could be a primary motor areas that serves as a final common pathway for many different behaviors, or a parietal area involved in visuomotor coordination. A straightforward way to obtain a functional degeneracy map is to perturb the activity of several areas using transcranial magnetic stimulation (TMS), to establish whether and to what extent this perturbation changes the activation of the target area. This can be done by combining TMS with imaging modalities such as PET, fMRI, or high-density EEG. Finally, degeneracy maps based on plastic connectivity are the set of brain areas whose influence on the outputs of a target brain area can be effectively potentiated or depressed. Here, too, TMS/neuroimaging approaches can be used to monitor changes in the efficacy of relevant pathways as a result of rehabilitation strategies (Perfetti et al., 2010; Price et al., 2010; Sarasso et al., 2010; Yourganov et al., 2010). On the basis of such degeneracy maps, it should be possible to design, implement and monitor rehabilitation procedures that are rationally based, tailored to the individual patient, and that require minimum patient effort.

Conclusion

This final contribution may leave the impression that the complexity of the brain is perhaps even more daunting from a theoretical perspective than from an empirical perspective. As shown by other articles in this issue, increasingly sophisticated experiments and clinical interventions are shedding light on brain function at multiple levels, from the effects of individual synapses to the behavior of an entire organism (McIntosh et al., 2010; Perfetti et al., 2010; Price et al., 2010; Protzner et al., 2010; Sarasso et al., 2010). By contrast, the theoretical approach outlined above makes it clear that as yet we cannot even begin to imagine, let alone chart exhaustively, the full set of integrated informational relationships generated by the corticothalamic system. Nevertheless, some of the theoretical implications may be of some relevance. One is that, while we may far from being able to describe the informational relationships the brain generates in all their richness, we are nevertheless able to experience them at every instant: they are consciousness itself. Another is that we should always pay attention not only to what the brain is doing, but also to what it could possibly do, which is vastly more. While this may begin as a theoretical prescription, it can acquire new practical meaning in the context of recovery of function.

Acknowledgments

I thank Chiara Cirelli, Lice Ghilardi, Frauke Harms, Christof Koch, Umberto Olcese and Puneet Rana for their help, and the McDonnell Foundation for support.

References

Alkire M.T., Hudetz A.G., Tononi G. Consciousness and anesthesia. *Science*, **322**: 876-880, 2008.

Bachmann T. *Microgenetic approach to the conscious mind*. Amsterdam, Philadelphia: John Benjamins Pub. Co, 2000.

Balduzzi D. and Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.*, **4**: e1000091, 2008.

Balduzzi D. and Tononi G. Qualia: the geometry of integrated information. *PLoS Comput. Biol.*, **5**: e1000462, 2009.

Bateson G. Steps to an ecology of mind; collected essays in anthropology, psychiatry, evolution, and epistemology. San Francisco, Chandler Pub. Co., 1972.

Block N. Two neural correlates of consciousness. *Trends. Cogn. Sci.*, **9**: 46-52, 2005.

Blumenfeld H. and Taylor J. Why do seizures cause loss of consciousness? *Neuroscientist*, **9**: 301-310, 2003.

Campbell D.T. 'Downward causation' in hierarchically organised biological systems. pp. 179-186. In: Dobzhansky F.J.A.T. (Ed.) *Studies in the philosophy of biology*, Berkeley, Los Angeles, University of California Press, 1974.

Cauller L. Layer I of primary sensory neocortex: where top-down converges upon bottom-up. *Behav. Brain Res.*, **71**: 163-170, 1995.

Cover T.M. and Thomas J.A. *Elements of information theory*. 2nd Edition. Hoboken, NJ, Wiley-Interscience, 2006.

Cowey A. and Walsh V. Magnetically induced phosphenes in sighted, blind and blindsighted observers. *Neuroreport*, **11**: 3269-3273, 2000.

Crick F. and Koch C. A framework for consciousness. *Nat. Neurosci.*, **6**: 119-126, 2003.

Edelman G.M. and Mountcastle V.B. The mindful brain: cortical organization and the group-selective theory of higher brain function. Cambridge, MIT Press, 1978.

Ferrarelli F., Massimini M., Sarasso S., Casali A., Riedner B.A., Angelini G., Tononi G., Pearce R.A. Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proc. Natl. Acad. Sci. USA*, **107**: 2681-2686, 2010.

Hudetz A.G. Suppressing consciousness: Mechanisms of general anesthesia. *Seminars in Anesthesia, Perioperative Medicine and Pain*, **25**: 196-204, 2006.

Imas O.A., Ropella K.M., Wood J.D., Hudetz A.G. Isoflurane disrupts anterio-posterior phase synchronization of flash-induced field potentials in the rat. *Neurosci. Lett.*, **402**: 216-221, 2006.

Jirsa V., Sporns O., Breakspear M., Deco G., McIntosh A.R. Towards the virtual brain: network modeling of the intact and the damaged brain. *Arch. Ital. Biol.*, **148**: **??**, 2010.

Koch C. The quest for consciousness: a neurobiological approach. Denver, CO., Roberts & Co., 2004.

Lamme V.A. Towards a true neural stance on consciousness. *Trends Cogn. Sci.*, **10**: 494-501, 2006.

Massimini M., Ferrarelli F., Murphy M.J., Huber R., Riedner B.A., Casarotto S., Tononi G. Cortical reactivity and effective connectivity during REM sleep in humans. *Cogn. Neurosci.*, 2010, in press.

McIntosh A.R. Towards a network theory of cognition. *Neural Netw.*, **13**: 861-870, 2000.

McIntosh A.R., Kovacevic N., Lippe S., Garrett D., Grady C., Jirsa V. The development of a noisy brain. *Arch. Ital. Biol.*, **148**: **??**, 2010.

Milner A.D., Goodale M.A. *The visual brain in action*. New York, Oxford University Press, 1995.

Moruzzi G. and Magoun H.W. Brain stem reticular formation and activation of the EEG. *Electroencephalog. Clin. Neurophysiol.*, **1**: 455-473, 1949.

Owen A.M., Schiff N.D., Laureys S. A new era of coma and consciousness science. *Prog. Brain Res.*, **177**: 399-411, 2009.

Pascual-Leone A. and Walsh V. Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, **292**: 510-512, 2001.

Perfetti B., Moisello C., Lanzafame S., Varanese S., Landsness E., Onofrj M., Di Rocco A., Tononi G., Ghilardi M.F. Attention modulation regulates both motor and non-motor performance: a high-density EEG study in Parkinson's disease. *Arch. Ital. Biol.*, **148**: **??**, 2010.

Pollen D.A. Fundamental requirements for primary visual perception. *Cereb. Cortex*, **18**: 1991-1998, 2008.

Pöppel E. and Artin T. *Mindworks: Time and conscious experience*. Boston, MA, US, Harcourt Brace Jovanovich, Inc., 1988.

Price C.J., Crinion J.T., Leff A.P., Richardson F.M., Schofield T., Prejawa S., Ramsden S., Gazarian K., Lawrence M., Ambridge L., Andric M., Small S.L., Seghier M.L. Lesion sites that predict the ability to gesture how an object is used. *Arch. Ital. Biol.*, **148**: **??**, 2010.

Protzner A.B., Valiante T.A., Kovacevic N., McCormick C., McAndrews M.P. Hippocampal signal complexity in mesial temporal lobe epilepsy: a noisy brain is a healthy brain. *Arch. Ital. Biol.*, **148**: **??**, 2010.

Sarasso S., Santhanam P., Maatta S., Poryiazova R., Ferrarelli F., Tononi G., Small S. Non-fluent aphasia and neural reorganization after speech therapy: insights from human sleep electrophysiology. *Arch. Ital. Biol.*, **148**: **??**, 2010.

Schiff N., Ribary U., Plum F., Llinas R. Words without mind. *J. Cogn. Neurosci.*, **11:** 650-656, 1999.

Schwabe L., Obermayer K., Angelucci A., Bressloff P.C. The role of feedback in shaping the extra-classical receptive field of cortical neurons: a recurrent network model. *J. Neurosci.*, **26**: 9117-9129, 2006.

Shannon C.E. and Weaver W. *The mathematical theory of communication*. Urbana, University of Illinois Press, 1963.

Silvanto J., Cowey A., Lavie N., Walsh V. Striate cortex (V1) activity gates awareness of motion. *Nat. Neurosci.*, **8**: 143-144, 2005.

Sperry R.W. Mind-brain interaction: mentalism, yes; dualism, no. *Neuroscience*, **5**: 195-206, 1980.

Szentagothai J. Downward causation? *Annu. Rev. Neurosci.*, **7**: 1-11, 1984.

Tononi G. Information measures for conscious experience. *Arch. Ital. Biol.,* **139**: 367-371, 2001.

Tononi G. An information integration theory of consciousness. *BMC Neurosci.*, **5**: 42, 2004.

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol. Bull.*, **215**: 216-242, 2008.

Tononi G. and Edelman G.M. Information: In the stimulus or in the context? *Behav. Brain Sci.*, **20**: 698-700, 1997.

Tononi G. and Laureys S. The Neurology of Consciousness: An Overview. pp. 375-412, In: Laureys S. and Tononi G. (Eds.) *The Neurology of Consciousness*, First Edition, Elsevier, 2009.

Tononi G., Sporns O., Edelman G.M. Reentry and the problem of integrating multiple cortical areas: simulation of dynamic integration in the visual system. *Cereb. Cortex*, **2**: 310-335, 1992.

Tononi G., Sporns O., Edelman G.M. A complexity measure for selective matching of signals by the brain. *Proc. Nat. Acad. Sci. USA*, **93**: 3422-3427, 1996.

Tononi G., Sporns O., Edelman G.M. Measures of degeneracy and redundancy in biological networks. *Proc. Nat. Acad. Sci. USA*, **96**: 3257-3262, 1999.

Tononi G., McIntosh A.R., Russell D.P., Edelman G.M. Functional clustering: Identifying strongly interactive brain regions in neuroimaging data. *Neuroimage*, **7**: 133-149, 1998.

Yourganov G., Schmah T., Small S.L., Rasmussen P.M., Strother P.M. Functional connectivity metrics during stroke recovery. *Arch. Ital. Biol.*, **148**:, 2010.

Notes

[1] In practice, one can think of such elementary mechanisms as universal logical gates, such as NOR gates. These are physically realizable, and if properly interconnected they can perform any computation, limited only by memory. Indeed, collections of such elements are equivalent to any particular Turing machine with a finite tape (NOR gates can also be used to implement memory; also, computations performed by neurons can be approximated by collections of logical gates). Note that with < 2 inputs, an element could not perform any integrating computations. With > 2 inputs, it could perform arbitrarily complex computations, but then it will have internal structure that would be reducible into simpler components. 2 inputs have 4 possible values, so there are $2^4 = 16$ ways of choosing an output based on the 4 inputs, i.e. 16 different mechanisms of input/output (truth) tables. Of these, NOR and NAND are the only ones that are a function of both inputs (integrative) and allow for negation. This is why all mechanisms can be reduced to a combination of NOR (or NAND) mechanisms. An element can have at most one self-connection, connecting it to its state the previous time step, thus implementing a form of memory by which the previous state of the element can influence its current state.

[2] This is a partial order of its connections K under the subset-relation ($\subset$, included in), which is represented by its Hasse diagram or lattice.

[3] Or one can evaluate the actual repertoire corresponding to the product of the actual repertoires generated independently by the parts (cf. Balduzzi and Tononi, 2008). One could also envision literally "cutting" those connections. Note also that the effective information generated by a mechanism in a state can be calculated not just with respect to a previous time step, but to a series of previous time steps. In other words, a certain mechanism may find itself in a certain state only if received a particular succession of inputs over time. In this case, the state of the mechanism specifies a sequence or 'melody'. Clearly, the brain must be richly endowed with such mechanisms.

[4] In the Shannon sense (Shannon and Weaver, 1963). The MIP is also conceptually related to the notion of max-flow min-cut in graph theory.

[5] In complex systems such as the brain connections are typically sparse, that is, their number grows much less than V^2. Since the number of subsets of K connections (power-set) grows as 2^K, it will typically be much lower than the number of partitions among V units (Bell number: the cross-over happens already at n = 5: n5 = 32, Bell = 52; for n = 10: n10 = 1024, Bell = 115975; the reason is that the power-set of connections only contains partitions of the form 123/4/5 or 12/3/4/5, and not of the form 12/34/5 or 123/45). Moreover, in the case of the brain one often has knowledge about plausible 'anatomical' bottlenecks, which are likely choices for the MIP.

[6] The need for normalization is evident if one considers a very asymmetric disconnection, say one where one part contains one unit and the other part all the other units. Assuming every unit has 2 inputs, this means that the lone unit can specify at most 2 bits/units across the partition (in the other part), and the other part can specify at most 1 bit/unit (the lone unit constituting the other part). Therefore, the best this particular partition can do is specify 3 bits, which is the normalization factor. By contrast, if the partition divides the system into subsets of similar size, or onto many different subsets, all or most inputs to each part can find 'unoc-

cupied' units in other parts, and the normalization factor is n or close to n.

[7] For example, consider a chain of units A->B->C, which generates 2 bits of effective information over one time step (B at t_1 specifies A at t_0 and C at t_1 specifies B at t_0). If one makes the output of unit B at t_0 independent of its computation at t_1, i.e. one forces temporal independence between B's role as a source and its role as a target, the system still generates 2 bits of effective information (B at t_1 still specifies A at t_0 and C at t_1 specifies B at t_0 even if B at t_1 and B at t_0 are made to be independent). Therefore, at one time step the temporal partition of the full system is its MIP, and φ for the chain is zero. Over two time steps, effective information for the system is 1 bit (unit C specifies unit A two time steps earlier). However, in this case making B's output independent of its inputs yields an effective information of zero bits, because C cannot specify A (since what B transmits at t_2 is made independent of what it computes at t_1). Thus, the chain is temporally integrated at 2 time steps, and its MIP yields 1 bit. This account implies that the identity of an element over time is established 'intrinsically' through causal interactions. For instance, in the case of the chain A->B->C it would remain undetermined at one time step that the target of one interaction (B at t_1) is the same as the source of another (B at t_0). That the interaction/information percolates through the two links of the chain, and thus the two links are indeed linked by a single element, would only be established at two time steps.

[8] Again, one should properly take into account temporal disconnections over a given intervals. Note that a complex for which $\varphi\ (X_{i,\,t1}) = ei(X_{i,\,t1})$ is called a complex *stricto sensu*, as it is strictly without parts.

[9] It is worth considering two different notions of complexes. The present notion enforces an '*exclusion principle*'. Such principle prescribes that a unit can belong to one and only one complex at any given time, which is the one having higher φ. According to this notion, any system of elements is informationally "crystallized" or condensed into sets of non-overlapping integrated complexes, each with higher φ values than its surroundings and higher or equal φ than its parts. Each complex has borders separating it from the outside and therefore from other complexes, although complexes can interact and thus exchange information. When the interactions among two or more complexes become strong enough that φ becomes higher for all their elements together than for each separate subset, there is a phase transition and the smaller complexes merge into a larger one. In this view, consciousness occurs exclusively over the local set of elements and at the spatial and temporal grain size at which φ reaches a maximum, and that set of elements can support only a single consciousness (that is, an individual consciousness is a local maximum of integrated information). A physical analogy for this notion of φ and complexes would be with the notion of cohesion: where cohesive forces are stronger than repulsive forces or external forces (including noise), a liquid or solid will form a "single entity", otherwise it will split along the fracture plane that is thermodynamically most likely (e.g. into droplets of water). The minimum information partition can be thought of as analogous to such a fracture plane, φ as analogous to the work required to separate the system along that partition, and a complex would be the set of elements having the maximum resistance to fracturing. Another, dynamical analogy would be with attractors: an element may be part of one of two or more competing attractors at any given time, unless the attractors are so strongly coupled that they merge. An alternative notion enforces a '*superposition*' principle. According to this notion (which was presented in some previous work), complexes can overlap, in whole or in part. Any set of elements with $\varphi > 0$ would then constitute a complex, even if it contains a subset of much higher φ. As a consequence, the same set of elements can support more than one consciousness, and different consciousnesses can overlap, although they share part of their informational structure. That an exclusion principle might apply is perhaps more in line with the intuitions that each of us has a single, sharply demarcated consciousness. Phenomena such as binocular rivalry and other kinds of metastable perception, not to mention dissociative personality disorders, are also suggestive of an exclusion principle. On the other hand, conventional neurophysiologic mechanisms might be sufficient to account for these phenomena, by "excluding" competing neuronal assemblies from a main complex of high φ and relegating them to complexes of much lower φ, which would be separately but only "dimly" aware. Endorsing a superposition principle requires accounting for how the informational structure (quale, see below) of a complex of high φ can be contained (or projected) into that of a larger complex of lower φ.

[10] Note that, in general, many subsets of connections will specify the same repertoire, and thus collapse onto the same point/q-arrow in Q. For example, in Fig. 4, the points specified by connections [2,4,6]

and connections [1,2,4,6] collapse. Asterisks indicate points generated by more than one subset of connections (only one subset is indicated).

[11] Under the partial order subset-relation of inclusion $\subset$.

[12] For a q-arrow joining a point in Q specified by a subset of connection m (a proper submechanism) and the point generated by adding subset r ($m \cup r$, another proper submechanism) to form a larger submechanism, its degree of information integration is defined as φ ($^{m}X_{i, t1} \parallel {}^{m \cup r}X_{i, t1}$). φ is positive if and only if the q-arrow cannot be decomposed into independent q-arrows in space and time (cf. the definition of entanglement in Balduzzi and Tononi, 2009).

[13] In both cases one is essentially applying Occam's razor, in fact literally so: entities should not be multiplied beyond necessity. In other words, there is no need to invoke an additional super-entity if it adds nothing to what smaller entities can do independently.

[14] Within a quale, one can define modes as subsets of q-arrows that are more densely integrated than surrounding q-arrows, which may underlie the modalities of experience (Tononi, 2008; Balduzzi and Tononi, 2009).

[15] For example, in the visual system the repertoire (point in Q and corresponding q-arrows) that specifies an invariant like "face", irrespective of its location, orientation, and various details, can only be generated after a given time interval by a chain of mechanisms converging onto units in high visual areas like IT. Thus, the quale containing that and similar repertoires will only unfold after a time interval sufficiently long for the relevant connections to be integrated within the same complex, and for the state of the relevant neurons to be specified with sufficient confidence (see note 18 below).

[16] In practice, one can start from the directed graph, and analyze the paths that link the vertices, with their length. To evaluate partitions and points in Q one would then consider exclusively connected components as potential complexes (if there are no paths between some vertices, they cannot be part of the same complex). Furthermore, one would only consider paths that are at most as long as the time interval being evaluated (if a subset of connections does not form a path in the allotted time, the corresponding submechanism cannot make a difference). Finally, within a given time interval, one would only consider multiple paths that have "joins" somewhere in the graph (if subsets of connections form paths that do not join in space, they cannot make a difference jointly).

[17] In other words, small φ, which is related to a measure of functional clustering introduced in Tononi et al., (1998), delineates complexes and their boundaries. Big PHI reflects instead the amount of information that is necessary to specify the shape of the quale, including al points in Q and their relative position. This information completely characterizes a particular experience in terms of its distinguishable relational features, thereby discriminating it from any other in the repertoire of possible experiences.

[18] The temporal grain size in a system of neurons may depend both on the time necessary to establish more and more confidently the state of the units – for example, that a neuron is ON or OFF based on a rate code, and on the time necessary for activity to percolate through the system due to conduction delays, synaptic delays, and processing delays (neural integration).

[19] Reflecting the duality or symmetry between causation and information, see above.

[20] On the other hand, what may be taken for a higher level of organization from the extrinsic perspective, may fail to generate high values of integrated information from the intrinsic perspective: only a few collections of macro-elements form a proper complex, even though they may be interacting.

[21] With probability density functions rather than repertoires, this would correspond to a volume.