LUMINOUS: Studying, Measuring and Altering Consciousness through information theory in the electrical brain

H2020-FETOPEN-2014-2015-RIA
Grant agreement No 686764

Work Package WP1: *Scientific framework*

Deliverable D1.1: Consciousness: models, metrics & intervention in the electric brain

Marcello Massimini, University of Milan (UMIL)
Silvia Casarotto, University of Milan (UMIL)
Angela Comanducci, University of Milan (UMIL)
Matteo Fecchio, University of Milan (UMIL)
Fabrice Wendling, Inserm (INS)
Siouar Bensaid, Inserm (INS)
Isabelle Merlet, Inserm (INS)
Pascal Benquet, Inserm (INS)
Julien Modolo, Inserm (INS)
Giulio Ruffini Starlab, (STARLAB)
Aureli Soria-Frisch, Starlab (STARLAB)
Eleni Kroupi, Starlab (STARLAB)
Josep Marco Pallares, Starlab, UB (STARLAB)
Marta Castellano, Starlab (STARLAB)

Release date: 28th November 2016
Status: *public*

Executive Summary

In this document we review three established theories of consciousness (Dynamic Core Hypothesis - DCH, Global Neuronal Workspace - GNW, Integrated Information Theory - IIT) and propose a new approach to consciousness based on algorithmic complexity (which we call K-theory - KT).

Then, we emphasize the translation of these theoretical insights into experimental methods and metrics of consciousness in order to provide a) guidance for the experimental work within the Luminous project and b) valuable information for developing computational brain models.

The guiding theme is the necessity of identifying and characterizing the most useful metrics of consciousness concerning the main objectives of the Luminous project, which include the observation and modulation of consciousness state in an open-loop as well as a closed-loop approach.

Deliverable Identification Sheet

IST Project No.	**H2020-FETOPEN-2014-2015-RIA Grant agreement No 686764**
Acronym	**LUMINOUS**
Full title	**Studying, Measuring and Altering Consciousness through information theory in the electrical brain**
Project URL	**http://www.luminous-project.eu/**
EU Project Officer	**Christiane Wilzeck**

Deliverable	**D1.1 Experimental plans M6**
Work package	**WP1: Scientific Framework**

Date of delivery	**Contractual**	M6: 16-AUg-2016	**Actual**	M9: 30 Nov-2016
Status	Version. 3.0		Final	
Nature	Report			
Dissemination Level	Public			

Authors (Partner)	Marcello Massimini (UMIL) Silvia Casarotto (UMIL) Angela Comanducci (UMIL) Fabrice Wendling, Inserm (INS) Siouar Bensaid (INS) Isabelle Merlet (INS) Pascal Benquet (INS) Julien Modolo (INS) Giulio Ruffini Starlab (STARLAB) Aureli Soria-Frisch (STARLAB) Eleni Kroupi (STARLAB) Josep Marco Pallares (STARLAB, UB) Marta Castellano (STARLAB)

Responsible Author	Marcello Massimini		**Email**	marcello.massimini@unimi.it
	Partner	UMIL	**Phone**	+39 02 50319885

Keywords	Consciousness, models, metrics, information, integration	

Version Log			
Issue Date	**Rev No.**	**Author**	**Change**
24-06-16	0	MM	Outline draft
30-05-16	1.0	FW,SB, IM	Edition + details in outline
19-08-16	2.1	GR, EK, JM, MC, ASF	Edition and addition of some sections (Starlab)
01-09-16	2.2	FW,SB, IM	Review and update
20-09-16	2.3	GR, EK, MC, ASF	Review and update
20-11-16	2.4	GR, EK, MC, ASF	Review and update
28-11-16	3.0	MM,SC,AC	Review and update, including reply to internal reviewers

Table of contents

1 Introduction

1.1 LUMINOUS and its primary objective

The LUMINOUS project aims at studying, measuring, and altering consciousness through the exploitation of information theoretic principles. Supported by computational neuroscience models, the project aims at creating non-invasive consciousness-probing technologies based on electroencephalography (EEG), magnetoencephalography (MEG) and peripheral and non-invasive electromagnetic brain stimulation (NIBS). Computational neuroscience models will be developed in parallel with human studies in sleep, anaesthesia, locked-in syndrome, disorders of consciousness, and in utero.

The general objectives of the project are:

O1. To provide a scientific, theoretical framework of consciousness for modelling, definition of metrics, and planning of experiments.

O2. To develop a physiologically validated computational neuroscience model to possibly derive the mechanisms of loss and recovery of consciousness and to predict the effects of NIBS on brain activity.

O3. To develop practical and clinically useful metrics for measuring consciousness in different physiological and pathological conditions and during non-invasive neuromodulation.

O4. To develop open- as well as closed-loop non-invasive technology to monitor and alter consciousness in research and medical applications.

O5. To validate and refine the measurement of consciousness in order to optimize diagnosis, prognosis, and treatment monitoring of patients with disorders of consciousness.

O6. To evaluate the ethical implications of these methodologies in view of their application in a clinical environment.

O7. To disseminate and exploit the project results via events, publications and, if applicable, protection of IPR.

1.2 The work of this deliverable

This deliverable provides an extensive review of the current literature about the theoretical models and experimental metrics of consciousness. In particular, it will describe the following four theories of consciousness (section 2): Dynamic Core Hypothesis (DCH), Global Neuronal Workspace (GNW), Integrated Information Theory (IIT) and Algorithmic Information (KT). These theoretical models will be discussed in connection with the principles of information and integration within the thalamocortical system (section 3.1) and in view of the development of computational neuroscience models (section 3.2). Section 4 will present a comprehensive list of the

available metrics of consciousness, organized into information-based, integration-based and information-integration-based metrics. For each metrics, a synthetic description of its computation is reported, together with its potential relevance for the Luminous project. The conclusion of this deliverable (Section 5) provides a summary of the guiding principles for the development of novel metrics of consciousness that can be inspired by current theoretical models and experimental work.

1.3 Abbreviations

ACC = anterior cingulate cortex; **ACE** = amplitude coalition entropy, **AUC** = area under the curve, **BOLD** = blood oxygen level dependent, **CA** = cellular automata, $\mathbf{C_d}$ = causal density, $\mathbf{C_N}$ = Neural Complexity, **CNV** = contingent negative variation, **CRS-R** = Coma Recovery Scale - Revised **CS** = conscious state, **CSD** = current source density, **DCH** = Dynamic Core Hypothesis, **DCM** = dynamic causal modeling, **DFA** = Detrended Fluctuation Analysis, **DLPFC** = Dorsolateral prefrontal cortex, **DOC** = disorders of consciousness, **EEG** = electroencephalography, **EMG** = electromyography, **EOG** = electrooculogram, **ERP** = event-related potential, **ERF** = event-related fields, **FFI** = feedforward inhibition, **fMRI**=functional magnetic resonance imaging, **GNW** = Global Neuronal Workspace, **GS** = generalized synchronization, **GW** = Global Workspace, **hd-EEG** = high density electroencephalography, **IIT** = integrated information theory of consciousness, **I/O** = input/output, **ImC** = imaginary coherence, **K** = Kolmogorov or Algorithmic Complexity, **KT** = K-theory of consciousness, **LDA** = linear discriminant analysis, **LFP** = local field potential, **LIS** = locked-in syndrome, **LOC** = loss of consciousness, **LORETA** = low-resolution brain electromagnetic tomography, **LSM** = liquid state machines, **LZW** = Lempel-Ziv-Welch compression algorithm, **MCS** = minimally conscious state, **MEG** = magnetoencephalography, **MDD** = major depression disorder, **MDL** = minimum description length, **MFDFA** = multifractal Detrended Fluctuation Analysis, **MI** = mutual information, **MML** = minimum message length, **MMN** = mismatch negativity, **MNE** = minimum norm estimate, **NCC** = neural correlates of consciousness, **NIBS** = non-invasive brain stimulation, **NLD** = normalized length density, **NN** = neural network, **NREM** = non rapid eye movement sleep, **PCI** = perturbational complexity index, **PDF** = probability density function, **PE** = permutation entropy, **PFC** = Prefrontal cortex, **PLV** = phase-locking value, **PP** = predictive processing theory, **PS** = phase synchronization, **PSC** = physical substrate of consciousness, **PV** = parvalbumine positive, **REM** = rapid eye movement sleep, **RNN** = recurrent neural network, **ROC** = receiver operating characteristic, **RTN** = reticular thalamic nucleus, **RTP** = rapid transition process, **sLORETA** = standardized low-resolution brain electromagnetic tomography, **SCE** = synchrony coalition entropy, **SOC** = self-organized criticality, **sPLV** = single-trial phase locking value, **SVM** = support vector machine, **SWS** = slow-wave sleep, **TC** = thalamo-cortical, **TMS** = transcranial magnetic stimulation, **TPO** = temporo-parietal-occipital junction, **US** = universal search, **UWS** = unresponsive wakefulness syndrome, **VAN** = visual awareness negativity, **VS** = vegetative state, **WLBMF** = Wavelet Leader Based Multifractal Formalism, **wMNE** = weighted minimum norm estimate, **wSMI** = weighted symbolic mutual information, **WT** = waiting times

2 Models of consciousness

Here we present the fundamental concepts underlying current theoretical models of consciousness and their potential implications for the implementation of empirical measures. While several proposals, ideas and hypotheses about the nature of consciousness and its relationships with the physical world have been put forward over the last the decades, we will restrict our focus to models that fulfill some basic criteria, which are in line with the fundamental objectives of the Luminous project. As outlined above, WP1 aims at i) providing foundations and methodology for the definition of experiments and measurements of consciousness and ii) defining a framework of consciousness for the implementation of a physiologically plausible computational model. Therefore, we will need to limit our analysis to theoretical models that (1) propose empirical methods and metrics and (2) refer to physiologically observable events that evolve on the spatial scale of neuronal assemblies and on the time scale of milliseconds. For example, we will not consider models relating consciousness to quantum mechanical phenomena in the brain because they are neither measurable nor they relate to any appreciable physiological dynamics. On the other hand, we will not consider models such as worldly discrimination or high-order thought theories because the measurements that they propose are purely behavioral (e.g. stimulus discriminability, confidence rating, post decision wagering) and do not explicitly relate to specific neuronal events.

With these criteria in mind, we will focus on four fundamental models: (1) the Dynamic Core Hypothesis (DCH), (2) the Global Neuronal Workspace (GNW), (3) the Integrated Information Theory (IIT) and (4) the K-theory of consciousness (KT) based on algorithmic information theory. These four models clearly differ in various respects but they all propose metrics of consciousness that can be related to neuronal events.

2.1 Dynamic Core Hypothesis

2.1.1 Context

The DCH was first introduced in a review paper by Giulio Tononi and Gerald Edelman in 1998 (Tononi and Edelman, 1998). At that time, the DCH immediately stood out because of two fundamental elements of novelty. First, instead of arguing whether a particular brain area or group of neurons contributes to consciousness or not, it focused on characterizing the kind of neural processes that may account for key properties of conscious experience. Second, it emphasized, for the first time explicitly, the informational aspects of consciousness and of the underlying processes.

2.1.2 Fundamental claim and definitions

The DCH starts from the consideration that conscious experience is, at once, integrated (each conscious scene is unified) and at the same time it is highly differentiated (within

a short time, one can experience any of a huge number of different conscious states). Thus, the neural process underlying conscious experience must be functionally integrated and, at the same time, highly differentiated. More specifically the DCH suggests the following:

a) a group of neurons can contribute directly to conscious experience only if it is part of a distributed functional cluster that achieves high integration in hundreds of milliseconds;

b) to sustain conscious experience, it is essential that this functional cluster is highly differentiated, as indicated by high values of complexity.

Since one of the aims of the present deliverable is to clarify and define a common framework, it is important to specify the terminology used in each model. In the context of DCH:

- The term *functional cluster* refers to a subset of a neural system with dynamics that display high statistical dependence internally and relatively lower dependence with elements outside the subset (Tononi et al., 1998a). This definition is important because it implies that the boundaries of the physical substrate of consciousness are continually shifting, with neuronal groups entering and exiting the core according to the flow of conscious contents.
- The term *reentry* refers to the recursive exchange of signals among brain areas through massively parallel reciprocal connections. Reentry is thought to bind the core together and must be distinguished from 'feed-back', which refers to the return of an error signal from an output to an input (Tononi et al., 1992).
- The term *complexity* has a very specific definition and provides a quantitative measure (Neural Complexity, C_N) of neural dynamics that is maximized by simultaneous high integration and high differentiation, as described in more detail in section 4.3.2 of the present deliverable.

In essence, a large *functional cluster* of neuronal groups together constitutes, through *reentry* on a time scale of hundreds of milliseconds, a unified neural process of high *complexity* termed the "dynamic core", in order to emphasize both its integration and its constantly changing activity patterns. The neuronal groups participating in the dynamic core select a global activity pattern within less than a second out of a very large repertoire and are much more strongly interactive among them than with the rest of the brain.

2.1.3 Relationships with neuroanatomy and neurophysiology

In light of the above, the DCH clearly predicts that some basic anatomical and neurophysiological properties are specifically relevant for consciousness:

A balanced anatomical structure. The first necessary condition is a pattern of structural connectivity characterized by a balance between functional segregation and functional integration. Evolutionary search approaches using genetic algorithms to specify the connection structure of simple network (32 nodes) showed that networks optimized for high complexity (as defined by DCH and measured as described in section 2.1) are characterized by high clustering coupled with a short characteristic path length (Sporns et al., 2000). Similarly, studies using neuronal networks showed that high regimes of complexity coincided with mixed connection patterns of both local and

long-range connections (Sporns and Tononi, 2002). Strikingly, these high complexity networks were very similar to the so-called 'small world' class of networks in which dense groups of nodes are connected by a small number of reciprocal links (Watts and Strogatz, 1998). These anatomical requirements are optimally matched by the general architecture of the thalamocortical system, which, according to DCH, is the necessary physical structure enabling the dynamic core.

Rapid reentrant interactions. Besides structural connectivity, the DCH identifies two fundamental aspects of neural activity that are characteristic of thalamocortical networks. The first is the presence of strong and rapid reentrant interactions among distributed sets of neurons. For example, this is evident in various kinds of cognitive tasks that are accompanied by the occurrence of short-term temporal correlations among distributed populations of neurons in the thalamocortical system (Singer and Gray, 1995), as well as by magnetoencephalographic study of binocular rivalry indicating that awareness of a stimulus is associated with increased coherence among distant brain regions (Tononi et al., 1998b).

Differentiated patterns of neuronal activity. Perhaps, the most interesting (and defining) claim of the DCH is that the strong and fast reentrant interactions among distributed groups of neurons are necessary but not sufficient for conscious experience. Thus, DCH predicts that if the repertoire of differentiated neural states is large, consciousness is possible. Conversely, if this repertoire is reduced, as when most groups of neurons in the cortex discharge synchronously and functional discriminations among them are obliterated, consciousness is curtailed or lost. This is strikingly demonstrated by the unconsciousness accompanying generalized seizures and slow-wave sleep. During generalized seizures and NREM sleep, the brain is not only active, but also most neurons fire in a highly synchronous manner indicating a loss of differentiation. In this perspective, the DCH suggests that the low-voltage, fast-activity EEG characteristic of waking and REM sleep reflects the availability of a rich and diverse repertoire of neural activity patterns.

2.1.4 Implication for empirical measures of consciousness

As detailed below (section 2.1), the DCH proposes a metric (C_N) that relates to physiologically observable variables such as balanced structural connectivity, fast reentrant interactions, and the differentiation of neuronal activity in thalamocortical networks. The fundamental insight offered by DCH and by the related measure C_N is that functional integration and functional differentiation need to be measured jointly. Thus, according to DCH, measuring the ignition of widespread activations or large scale synchronous dynamics would not suffice *per se*. Indeed, one would not know whether these integrated dynamics are differentiated or stereotypical (such as the ones recorded during unconscious seizures or NREM sleep). On the other hand, simply measuring the algorithmic complexity or the entropy of ongoing time series would not do. In this case, in fact, one would not know whether these differentiated patterns are generated by one system of interacting elements or by a collection of independent elements. In this perspective, the DCH highlights a fundamental principle that should inspire any brain-based measure of consciousness. As caveats, it is important to note that C_N in its full form is computationally prohibitive to calculate on large networks and that nor the DCH neither C_N take into account causal (directed) interactions, but only temporal correlations (at zero-lag).

2.2 Global Neuronal Workspace

2.2.1 Context

The Global Neuronal Workspace (GNW) has been developed by Bernard Baars (Baars, 1988), and expended upon and linked with neural correlates by Stanislas Dehaene et al., resulting in the Global Neuronal Workspace theory (Dehaene et al., 1998, 2003).

2.2.2 Fundamental claim and definitions

GNW postulates that conscious information is globally available within the brain (*Figure 2-1*), and that two different computational systems co-exist: 1) distributed, local "processors" or circuits operating in parallel throughout the brain and being specialized ("unconscious"); and 2) a "global" workspace (GW) formed of a distributed network connected to other cortical areas, involved in conscious perception (Baars, 1988). It has been suggested that these horizontal cortico-cortical connections are glutamatergic, and that these projections enable a possible "ignition" (Dehaene et al., 2003), defined as a sustained, brain-scale activation following a stimulus. According to GNW, it is only through ignition that one can gain conscious perception of a stimulus: if the stimulus only evokes a localized, short-duration brain response, no ignition occurs and conscious perception is not possible. Conversely, through activation of a brain-scale network, conscious experience can occur. The GW constitutes a sort of "bottleneck", in that only a fraction of inputs can access it: one can imagine that, at a given moment, many possible inputs are competing to access to the GW. The inputs able to access consciousness depend on several factors, including selective attention, which will favor the perception of specific stimuli. During GW activity, it is suggested that only a fraction of neurons are activated, while the remaining GW neurons are inhibited through local, horizontal inhibitory pathways.

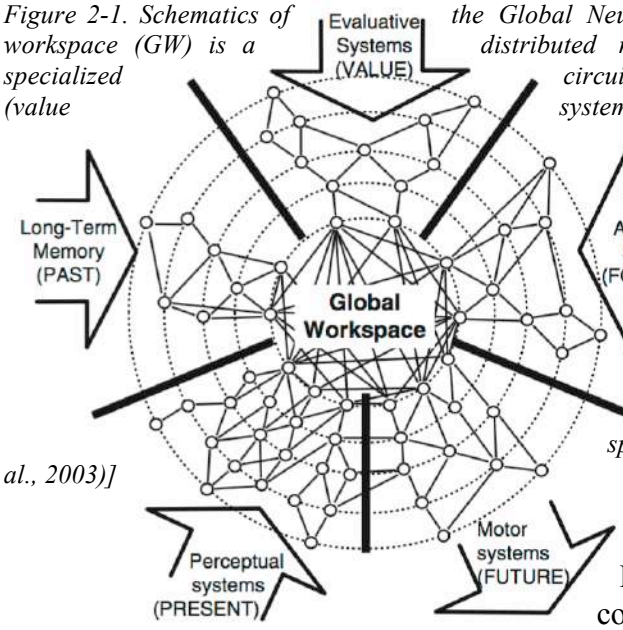

Figure 2-1. Schematics of the Global Neuronal Workspace theory (GNW). The global workspace (GW) is a distributed network associating functionally segregated specialized circuits, involved in perception, memory, evaluation (value system), which together can be used to determine motor outputs. Critically, the GW also involves the (selective) attention system, which corresponds to the "awareness" component of consciousness. The "wakefulness" component of consciousness is not directly present, since there is no dedicated circuit for wakefulness as there is for attention: wakefulness is one of the possible dynamical states of the brain and not a specialized circuit. [Adapted from (Dehaene et al., 2003)]

Importantly, the GW does not reduce consciousness to working memory (WM)

only, as highlighted by Baars (Baars, 1988). Instead, a conscious system operating as in GNW is able to account for the functions usually attributed to WM (Baars, 1988). In GW, the numerous processors operating in parallel contribute to various functions, and do not reach consciousness ("unconscious"), while the brain regions involved in the global workspace work in a coordinated manner, integrating information from the "unconscious" processors, in order to form a conscious perception.

2.2.3 Relationships with neuroanatomy and neurophysiology

While GW theory in its original formulation was mostly based on general principles (Baars, 1988), subsequent research efforts have attempted to link GNW principles with brain neuroanatomy and neurophysiology. Later studies investigating GW neural correlates have proposed that neurons supposedly related to the GW are indeed cortical pyramidal neurons from layers 2 and 3, which are involved in associating efferents and afferents (Dehaene et al., 1998). Interestingly, these pyramidal cells have also significant vertical, bidirectional connectivity with thalamic nuclei, through layer 5 neurons. Therefore, this subset of heavily horizontally connected pyramidal neurons, distributed among different brain regions, could enable brain-scale activation following a stimulus, a pre-requirement for conscious perception. At the local level, top-down connectivity is involved in recruiting specialized processors (unconscious processes).

Importantly, it is hypothesized that GW activation can be modulated, for example by mesencephalic reticular neurons (Dehaene et al., 1998), which control the level of depolarization of thalamocortical cells and therefore the pattern of thalamocortical activity: either slow, stereotyped, synchronized thalamocortical activity (e.g., sleep), or fast, weakly synchronized thalamocortical activity (e.g., awake state). In a previous work on GNW, the authors proposed the dorsolateral prefrontal cortex (DLPFC) and anterior cingulate cortex (ACC) as significantly involved regions. However, in a recent review paper (Koch et al., 2016), lobotomy patients have been cited to show that the frontal lobe is not crucial to achieve consciousness, challenging the notion that the DLPFC is involved in the GW. According to (Aru et al., 2012), the PFC might be rather involved in the maintenance of conscious perception rather than in the generation of conscious perception itself.

In an attempt to implement GW in a detailed computational model of large-scale brain activity while taking into account single-cell activity and realistic synaptic dynamics, Dehaene et al. (Dehaene et al., 2003) modelled the "attentional blink" phenomenon (impairment in stimulus perception if it is presented 100-500 ms after another stimulus). The model implemented i) AMPA, NMDA and GABA synaptic currents, ii) top-down vertical connectivity through cortical layers, and iii) horizontal connectivity through glutamatergic cortico-cortical projections. The model also included i) neuromodulatory input (e.g., from the brainstem), ii) a columnar organization and connectivity with horizontal, intra-column inhibition, and iii) transmission delays that were a function of distance. The model architecture included, following the initial principles formulated by Baars (Baars, 1988) numerous interacting thalamocortical columns. This simplified GW computational model, featuring neuroanatomical (corticocortical connectivity) and neurophysiological (synaptic currents with specific kinetics, neuromodulatory inputs) knowledge, accounted for the experimental data acquired in healthy volunteers, providing further support for the GNW.

2.2.4 Implication for empirical measures of consciousness

The GW computational model developed by Dehaene et al. (Dehaene et al., 2003) proposed that the P300 component of event-related potentials (ERP) is directly linked to sudden and global activation of workspace neurons. However, this notion has been challenged by Aru et al. (Aru et al., 2012), suggesting that instead of a single NCC (neural correlate of consciousness) network, three distinct NCC should be considered: the NCC for conscious perception itself, and pre- and post-NCC networks involved in pre- and post- conscious experience. The authors argue that, since the P300 can be abolished if the participant keeps in working memory the stimulus before its presentation, the P300 could be the marker of a transfer from the NCC network following a conscious experience to working memory. P300 would be therefore a marker of post-NCC networks activation, rather than a single NCC of the level of consciousness. Therefore, the P300 component in itself does not appear as an appropriate measure of consciousness in the context of GNW.

The GW also involves an "ignition" phenomenon that can bring a stimulus towards a conscious perception, beyond the initial, purely sensory response. In this respect, a fundamental prediction of GW is that a large-scale, distributed network is involved in conscious perception. If the evoked response to a stimulus is too local and/or weak, the stimulus will not reach consciousness. Therefore, it might be appropriate to suggest a measure of functional connectivity between distant brain regions to characterize the response to a given stimulus, possibly including, but not limited to, the total length of the graph representing the involved network, the mean clustering degree, the total number of nodes (discussed in *Section 4.2*). Therefore, since GNW has strong assumptions on what differentiates a conscious/unconscious experience in terms of brain networks, it should be possible to use measures from graph theory to provide candidate measures of consciousness.

While measures from graph theory would provide information about ignition and brain-scale coordination (characteristic of GW), this would not provide local information about neuronal dynamics (e.g. EEG activity). As discussed above, the use of P300 would not be appropriate. However, in terms of fine temporal activity linked with conscious perception in GNW, Dehaene et al. (Dehaene et al., 2003)link distant synchronization in the gamma band with the occurrence of a conscious perception, since gamma synchronization drops during the attentional blink (in the model and in experiments). Since there are consciousness states in which gamma synchronization is present while consciousness is impaired (e.g. during seizures or anesthesia, see (Koch et al., 2016)), we can state that gamma synchronization in itself is a necessary, but not sufficient condition for consciousness. Overall, from the predictions of GNW, it is possible to envision experimentally testable, measurable quantities that would be related to the state of consciousness. Since network properties and gamma band synchronization, which are involved in GNW, do not appear to provide reliable consciousness measures on their own, one might envision to combine them into a unique quantitative measure of consciousness, which is discussed in *Section 4.2*.

2.3 Integrated Information Theory

2.3.1 Context

The Integrated Information Theory of consciousness (IIT) has been developed by Giulio Tononi over the past several years (Oizumi et al., 2014; Tononi, 2004, 2008, 2012, 2015; Tononi and Koch, 2015) as an evolution of the Dynamic Core Hypothesis initially formulated in 1998 and described above. Not surprisingly, there is broad convergence between the two models on key points, such that consciousness is associated with pattern of differentiated neural activity in distributed thalamocortical circuits and that integration through reentrant interactions among them are important. However, IIT goes beyond by generalizing the principles underlying DCH in order to establish a principled way to assess consciousness in the physical world. In this respect, IIT differs from all other models, which are fundamentally agnostic about the relationship of consciousness to physical substrates markedly different from the brain, as for example, digital computers. In order to do this, IIT addresses "the hard problem" in a new way: it does not start from the brain and asks how it could give rise to experience; instead it starts from the *axioms***,** or essential phenomenal properties of experience and infers *postulates* about the characteristics that are required from its physical substrate. Based on this, IIT then presents a mathematical framework for evaluating the quality and quantity of consciousness (Oizumi et al., 2014; Tononi, 2012, 2015; Tononi and Koch, 2015).

2.3.2 Fundamental claim and definitions

Below we provide a summary of the axioms of conscious experience and corresponding postulates of IIT, and show how they can be used, in principle, to identify the physical substrate of consciousness (PSC).

The first axiom of IIT states that *experience exists intrinsically*. As recognized by Descartes (Descartes, 1999), one's own experience is the only thing whose existence is immediately and absolutely evident, and it exists for oneself, from one's own intrinsic perspective. The corresponding postulate states that the PSC must also exist intrinsically. For something to exist in a physical sense, it must have cause–effect power; that is, it must be possible to make a difference to it (for example by changing its state), and it must be able to make a difference to something. Moreover, the PSC must exist intrinsically; that is, it must have cause-effect power for itself, from its own intrinsic perspective. A neuron in the brain, for example, satisfies the criterion for existence, as it has two or more internal states (such as active and inactive) that can be affected by inputs (causes) and its output can make a difference to other neurons (effects). Furthermore, a minimal system consisting of two interconnected neurons satisfies the criterion of intrinsic existence since, through their reciprocal interactions, the system can make a difference to itself.

The axiom of composition states that *experience is structured*, as it is composed of phenomenological distinctions that exist within it. For example, in an experience, one may distinguish a piano, a blue color, a book, countless spatial locations, and so on. Based on this axiom, IIT postulates that the elements that constitute the PSC must also

have cause–effect power within the PSC, either alone or in combination (that compose first- and higher-order mechanisms, respectively).

The axiom of information states that *experience is specific*, being composed of a particular set of phenomenal distinctions (qualia), which make it what it is and different from other experiences. For example, the content of one's current experience might be composed of seeing a book (rather than seeing no book), which is blue (rather than not blue), and so on for all other possible contents of consciousness. The corresponding postulate states that the PSC specifies a cause–effect structure of a specific form, which makes it different from other possible ones. A cause–effect structure is defined as the set of cause–effect repertoires specified by all the mechanisms of a system. A cause–effect repertoire specifies how a mechanism in its current state affects the probability distribution of past and future states of the system.

The axiom of integration states that *experience is unitary*, meaning that it is composed of a set of phenomenal distinctions, bound together in various ways, that is irreducible to non-interdependent subsets. For example, one experiences a whole visual scene and that experience cannot be subdivided into independent experiences of the left and right sides of the visual field. In other words, the content of an experience (information) is integrated within a unitary consciousness**.** The corresponding postulate states that the cause–effect structure specified by the PSC must also be unitary; that is, it must be irreducible to the cause-effect structure specified by non-interdependent sub-systems. Note that, from the intrinsic perspective of the system, integration requires that every part of the system has both causes and effects within the rest of the system, which implies bidirectional interactions. The irreducibility of a conceptual structure is measured as integrated information (denoted Φ, the minimal distance between an intact and a partitioned cause-effect structure). The integration postulate also requires the irreducibility of each cause-effect repertoire (denoted φ, the minimal distance between an intact and a partitioned cause-effect repertoire) and the irreducibility of relations between overlapping cause-effect repertoires.

The axiom of exclusion states that an *experience is definite* in its content and spatio-temporal grain. For example, the content of one's present experience includes seeing one's hands on the piano, the books on the piano, one of which is blue, and so on, but one is not having an experience with less content (for example, the same scene in black and white, lacking the phenomenal distinction between colored and not colored) or with more content (for example, including the additional phenomenal distinction of feeling one's blood pressure as high or low). The duration of the instant of consciousness is also definite, ranging from a few tens of ms to a few hundred ms, rather than lasting a few ms or a few minutes (Bachmann, 2000; Holcombe, 2009; Pöppel, 1988). The corresponding postulate states that the cause-effect structure specified by the PSC must also be definite: it must specify a definite set of cause-effect repertoires over a definite set of elements, neither less nor more, at a definite spatio-temporal grain, neither finer nor coarser. Since a prerequisite for intrinsic existence is having irreducible cause-effect power, the cause-effect structure that actually exists over a set of elements and spatio-temporal grains is the one that is maximally irreducible (Φ^{max}), called a conceptual structure. As a consequence, any cause-effect structure overlapping over the same set of elements and spatio-temporal grain is excluded. The exclusion postulate also requires the maximal irreducibility of cause-effect repertoires (denoted φ^{max}), called concepts, and of relations between overlapping concepts.

On these bases, IIT formulates its central claim as follows: *an experience is identical to a conceptual structure* that is *maximally irreducible intrinsically*. Hence, a conceptual structure completely specifies both the quantity and the quality of experience. Specifically, *how much* the system exists—the quantity or level of consciousness—is measured by its Φ^{value} (as described in section 4.3.3), which quantifies the intrinsic irreducibility of the conceptual structure. If a system has $\Phi = 0$, meaning that its cause–effect power is completely reducible to that of its parts, it cannot lay claim to existing. If $\Phi > 0$, the system cannot be reduced to its parts, so it exists in and of itself. More generally, the larger Φ, the more a system can lay claim to existing in a fuller sense than systems with lower Φ. It is important to note that, according to IIT, the postulated identity is between an experience and the conceptual structure specified by the PSC, not between an experience and the set of elements in a state constituting the PSC.

2.3.3 Relationships with neuroanatomy and neurophysiology

IIT provides a principled explanation for several seemingly disparate facts about the PSC. For example, IIT can explain why the cerebral cortex is important for consciousness but the cerebellum is not. In general, the coexistence of functional specialization and integration in the cerebral cortex is ideally suited to integrating information. Specifically, the grid-like horizontal connectivity among neurons in topographically organized areas in posterior cortex, augmented by converging–diverging vertical connectivity linking neurons along sensory hierarchies, should yield high values of Φ. By contrast, cerebellar microzones that process inputs and produce outputs that are feedforward and largely independent of each other cannot form a large complex; nor can they be incorporated into a cortical high Φ complex even though each cerebellar microzone may be functionally connected with a portion of the cerebral cortex (Oizumi et al., 2014). In principle, these differences in organization can explain why lesions of the cerebellum, which has four times more neurons than the cerebral cortex (Herculano-Houzel, 2012), do not seem to affect consciousness (Lemon and Edgley, 2010; Yu et al., 2015). Furthermore, circuits providing inputs and outputs to a major complex may not contribute to consciousness directly. This seems to be the case with neural activity in peripheral sensory and motor pathways, as well as within circuits looping out and back into the cortex through the basal ganglia (Caparros-Lefebvre et al., 1997; Jain et al., 2013; Straussberg et al., 2002), despite their manifest ability to affect cortical activity and thereby to influence the content of experience indirectly.

IIT also accounts for the fading of consciousness during slow-wave sleep, when cortical neurons fire but, owing to changes in neuromodulation, they become bistable — that is, any input quickly triggers a stereotypical neuronal down state, after which neurons enter an up state and activity resumes stochastically (Steriade et al., 2001). Bistability implies a generalized loss of both selectivity (causal convergence or degeneracy) and effectiveness (causal divergence or indeterminism) (Hoel et al., 2013) that result in a breakdown of the cause-effect structure and thus of integrated information. Findings from a recent study that used intracranial stimulation and recordings in individuals with epilepsy are consistent with this account (Pigorini et al., 2015). During wakefulness, electrical stimulation of the cortex triggered a chain of deterministic phase-locked activations, whereas during slow wave sleep the same input induced a stereotyped slow wave that was associated with a cortical down-state (that is, a suppression of power >20 Hz). After the down-state cortical activity resumed to wakefulness-like levels, but the phase-locking to the stimulus was lost, indicative of a break in the cause–effect

repertoire. Similar considerations would explain why integrated information is impaired when consciousness fades despite the increased level of activity and synchronization that occurs early on during generalized seizures (Blumenfeld, 2012).

2.3.4 Implication for empirical measures of consciousness

As described in section 4.3.3, Φ, just like C_N, cannot be measured for any nontrivial real-world systems. However, IIT and the related measure Φ, represent an evolution and a generalization of C_N and provides additional principles that are important for the development of empirically tractable measures of consciousness. Crucially, according to IIT integrated information must be measured by a causal and intrinsic perspective. Hence, unlike DCH and GNW, for IIT it is not enough to observe pattern of temporal correlations from an extrinsic perspective. In order to assess the capacity for consciousness one needs to perturb the system directly to measure the amount of irreducible information that can be generated through intrinsic causal interactions. More in general and more relevant for the Luminous project, it is important to note that the term information is used very differently in IIT and in Shannon's theory of communication (Oizumi et al., 2014). In IIT information is causal and intrinsic: it is assessed from the intrinsic perspective of a system based on perturbations of its internal states (cause-effect power). Crucially, in IIT information must be integrated. This means that if partitioning a system makes no difference to it, there is no system to begin with. By contrast, Shannon information is observational and extrinsic—it is assessed from the extrinsic perspective of an observer and it quantifies how accurately input signals can be decoded from the output signals transmitted across a noisy channel. Crucially, it does not require integration (Oizumi et al., 2014). In essence, IIT claims that what matters for consciousness is the amount of information that a system can integrate through internal causal interaction and implies that this quantity can only be measured by a perturbational perspective. An empirical measure, the perturbational complexity index (PCI), was recently introduced as a practical proxy for Φ^{max}. Calculating PCI involves two steps: (i) perturbing a subset of cortical neurons with transcranial magnetic stimulation (TMS, "zapping") to engage distributed, deterministic interactions in the brain (integration) and (ii) measuring the incompressibility (algorithmic complexity, "zipping") of the resulting electrocortical responses (information). PCI is high only if brain responses are both integrated and differentiated, corresponding to a distributed spatiotemporal pattern of deterministic activations that is complex, hence not very compressible. Thus, PCI indexes something that is very different from the ignition proposed by GNW, as it would be low in case of widespread, stable response such as the P3b. In line with IIT, PCI thus gauges the amount of information that is generated through causal interactions intrinsic to the thalamocortical system.

2.4 Algorithmic information theory of consciousness (KT)

2.4.1 Context

K-theory (or KT) is a theory under development. Its name refers to its roots in Kolmogorov (or Algorithmic) complexity, and originates from earlier studies of the relation between cognition, physics and algorithmic information theory (Ruffini, 2007, 2009). What follows is an abridged version of Ruffini (2016, *submitted*). KT is closely related to IIT and Predictive Coding, which are briefly discussed below. The main difference is the focus on algorithmic aspects of information.

2.4.2 Fundamental claims and definitions:

KT starts from the concept of cognition in the context of algorithmic information and probability theory, following earlier work (Ruffini, 2007, 2009). The central idea is that brains strive to model their input/output fluxes of information (I/O) — with *simplicity* as a fundamental driving principle. It is important to advance here that the definition of brain is not limited to the human cortex, for example. As we discuss below, brains, agents or cognitive systems are to be identified with complex patterns embedded in a mathematical structure with certain properties (e.g., enabling computation and compression). This definition may apply to the entire human body (and other potential technology mediated future appendages), for example — such details are not crucial at this point. However, consciousness research does show that certain body parts are not needed to leave observable records and reports of the experience of consciousness (e.g., the cerebellum (Tononi and Koch, 2015)), while others are crucial (posterior cortex). Whether these records are necessary for consciousness itself is of course a different matter. What is becoming increasingly clear is that the brain is involved—to some extent— globally when conscious phenomena are experienced (Godwin et al., 2015).

A summary of what we may call the K-theory of cognition and consciousness is a follows. We first start with the subjective view (my brain and my conscious experience):

1) There is information and I am conscious (Ruffini, 2007, 2009). This is my subjective experience, an analog of the cartesian angular stone. Information here refers to the messages/signals traveling in and out of my brain or even within parts of my brain (**I/O streams**), and to Shannon's definition of the information conveyed by those messages. Consciousness refers to my conscious *experience*, e.g., of the feel of wind in my face during a sunny day in the beach (hopefully the reader is also conscious and can relate to this experience).

2) "Reality" is a model my brain has built and continues to refine based on input-output information. Brains are model builders, compressors of information for survival. Cognition is seen as equivalent to modeling and compression.

Then we shift to the objective view: what kind of mathematical structures could give rise to the above?

3) Essentially, universal Turing machines are needed. To explore this further, we can consider cellular automata (CA) models of the universe as a concrete example for

the formalization of the concepts of information transfer and computation, including the "embedding" of Turing machines, and of "brains" as special complex patterns that can actually represent (model) parts of the universe. CAs can instantiate, as part of their computation, sub-Turing machines (reversible or not) producing "complex" data chatter, complex behavior, in a precise mathematical sense. We point to existing connections between algorithmically generated complexity, compressive complexity and other measures of complexity (e.g., multiscale entropy, fluctuation analysis, fractal structure, complex networks, avalanche analysis, etc., which can be studied in models such as CA. Recurrent Neural Networks are also known to be Turing complete and provide a paradigm to study these connections closer to "wetware".

4) We return to the subjective and hypothesize that (graded, multidimensional) consciousness arises or is at least shaped, somehow, from the existence and updating of such complex patterns. In this sense, we are very close to Integration Information Theory (further discussed below).

Finally, based on the prior items and shifting to empirical application, KT proposes some methods to characterize complex behavior (internal / physiological or external) based on algorithmic complexity theory, and lines for further research to relate Kolmogorov complexity (K) to other "flavors" of complexity.

The framework is based on a series of definitions and hypothesis, briefly summarized here:

Definition 1. A model is an algorithm that generates a data set efficiently (equivalently, a compressor), i.e., succinctly and with minimal error.

For the purposes here, we can think of a model as a program running in a computer, such as CA or, more to the point, a recurrent neural network (RNN, which is also Turing complete framework). To illustrate this, let us begin with a simple type, a feed-forward network such as the ones used for image classification. Such a network encodes a function h(x). Let us imagine a function h(x) such that when fed an image of hand x it outputs 1, and 0 otherwise. In mathematical language we would say that this function is an invariant over the manifold of hands. If this function is actually a model, can we use it to generate images of hands? Yes: find the set of points $H = \{x \mid f(x) = 1\}$, and list them (in general, there will be an infinitely many solutions, which brings to light the meaning of compression). There will be many such points, and we need some means to enumerate them. Now we see that we can use h(x) and a parameter to select an element in the list to unpack the function into an image of a concrete hand. A trained neural network (NN) packages all the images it has seen before in this way. In this way, we can talk about the models encoded by neural networks.

Definition 2. A cognitive system or agent (or brain) is a model-building semi-isolated computational system capable of controlling some of its couplings/information interfaces with the rest of the universe driven by an internal optimization function.

In *Figure 2-2* we point to some key elements in a generic model and agent. The first is comprised by the model and error stream. Together they determine the model success given a data stream. Model and error stream – both of which are essential and require feedback for learning – are passed onto an action module that makes decisions and sends outputs streams, which are also form a feedback loop. The better the fit and the power of the model (e.g., integration of multi sensory streams), the stronger the experience (how real it feels). The model itself is a multidimensional object, and can easily account for the variety of experiences. The objective function can be seen as

assigning valence, positive, negative or null. We also note that evolution and natural selection will drive strongly the design optimization functions (e.g., homeostasis). Finally, we do not explain how the model has been discovered, only that one is available, perhaps assembled from available ones by integrating and/or tweaking them, and that it is malleable (learning can occur).

The details of model building and compression within the KT framework is reported below (*Appendix A*
KT: model building and compression).

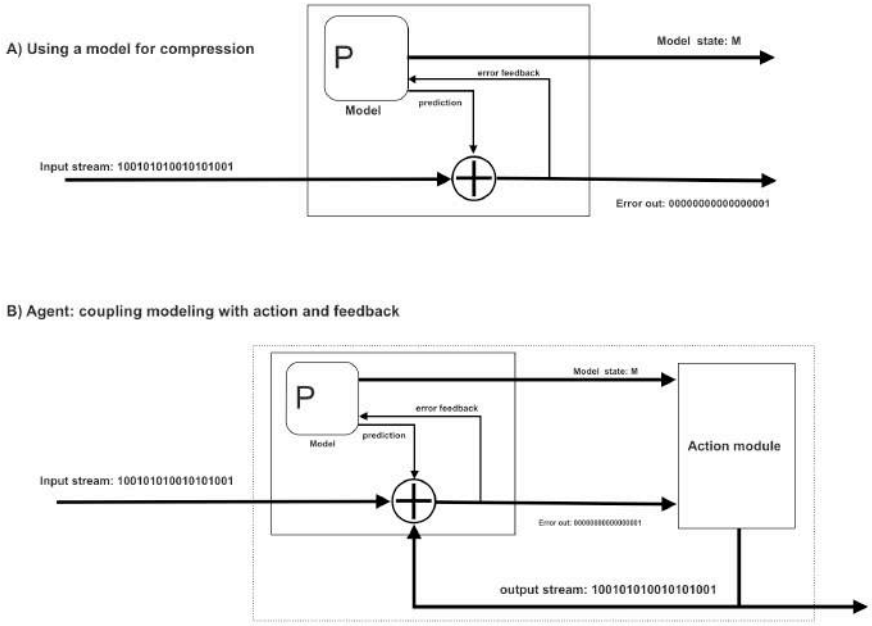

Figure 2-2: Top: Example of using a model for predictive compression. Bottom: An agent with coupled modeling and action modules via feedback (the action module must include an optimization function).

In essence, KT postulates that conscious experience emerges or is at least shaped in cognitive systems by the act of compression of information using successful models. Insofar we declare consciousness to represent the act of tracking and modeling such input/output streams, metrics for consciousness can be derived from the apparent complexity of brain data and its mutual information with external inputs or its own outputs. Further, we make two predictions: consciousness level will be a (possibly non-monotonic) function of the apparent complexity of brain activity data, and b) brain activity data will reflect high mutual information (MI) with external data (I/O) when available (as perturbed by sensory or possibly non-invasive brain stimulation, NIBS) in proportion to awareness.

2.4.3 Relation to other information based theories of consciousness

KT is closely related to other important theories of consciousness, and it may actually provide a conceptual link among them. The first point of contact is clearly with Predictive Processing theory. **Predictive processing theory (PP**) (Clark, 2013; Hohwy, 2013; Seth, 2013, 2014) is closely related to KT, but focuses on the predictive part afforded by the efficient description of I/Os. It maintains that in order to support

adaptation, the brain must discover information about the likely external "causes" of sensory signals using only information in the flux of the sensory signals them- selves. According to PP, perception solves this problem via probabilistic, knowledge-driven inference on the causes of sensory signals. The main differences are that KT emphasizes the roles of both input and output streams and that the "causes" are defined by models which are derived by the objective of compressing information (which include Bayesian models). Nevertheless, the parallels in KT and PP are strong.

In **Integration Information theory** (IIT), the most important property of consciousness is that it is "extraordinarily informative" (Tononi and Koch, 2008). It maintains that when we experience a conscious state, this rules out a huge number of possibilities. Here we can see that KT has links with this view. If reality is experienced as a simple model, our belief in this particular model "now" (driven by the input/output streams up to this moment) rules out — or lowers our belief — in a very efficient manner all other models and states for the experienced information streams. The simpler the model, the stronger will also be our belief in it (as encapsulated in Occam's razor or minimum description length, MDL). So with regard to IIT, in KT we make a transition from "information" to "algorithmic information". IIT also emphasizes that information associated to a conscious state must be integrated information: the conscious state is an integrated whole that cannot be divided into sub-experiences, data from the input/output streams must be closely bound together. KT provides here an analogous mechanism to "weave" together information: a good model will by definition integrate available information streams into a coherent whole (in an algorithm or model). An RNN instantiating a simple model should display integration. While IIT states that the level of consciousness of a physical system is related to the repertoire of causal states (information) available to the system as a whole (integration), here we would say, perhaps in simpler terms, that the level of consciousness of a physical system is related to its capability to model its input/output information streams in an efficient manner. Economy of description implies both a vast repertoire (reduction of uncertainty or information) and integration of information.

We note the statement above: "In essence, IIT claims that what matters for consciousness is the amount of information that a system can integrate through internal causal interaction and implies that this quantity can only be measured by a perturbative perspective." In KT, we would simply say that what matters in consciousness is the existence and running of simple, integrative models of data streams. This should in principle be measurable perturbatively or not. An important difference between IIT and KT is the use of algorithmic information in the latter. Shannon entropy metrics cannot capture all aspects of algorithmic information.

Global Neuronal Workspace theory (GNW) (Baars, 1988; Dehaene et al., 2003) has common elements with IIT although the focus shifts further to the "hardware" involved and to the analysis of experimental scenarios in which it has been shown that consciousness emerges when the brain works as a whole to process information. From the viewpoint of KT, the integral modeling system we call a brain is better equipped to integrate and compress information than its parts. The fact that effective real-time models may require parallel information processing to run is probably related to the fact that the entire brain is involved in those conscious moments. Since the original work of Baars, Dehaene and others (see, e.g., the recent results in (Godwin et al., 2015) have identified global brain events to correspond to the conscious experience in numerous experiments. Dehaene et al. (Dehaene et al., 2014) argues that "When we say that we are aware of a certain piece of information, what we mean is just this: the information

has entered into a specific storage area that makes it available to the rest of the brain". Moreover, "The flexible dissemination of information, ... is a characteristic property of the conscious state".

2.4.4 Relationships with neuroanatomy and neurophysiology

From the point of view of KT, consciousness is associated to the successful modeling events. Crucially, such events require integration of information from a variety of sensory and effective systems that must bring information together for model validation. Data must thus flow from a variety of sub-systems involving separate brain areas, but it needs to come together as well for integrative error checking against a running model. A candidate for such a location could be the temporo-parietal-occipital junction (TPO), an association area that integrates information from auditory, visual and somatosensory information, as well as from the thalamus and limbic system. It has been identified as a candidate for the generation of Presence qualia and as a full neural correlate of consciousness (Koch et al., 2016). There may be other locations where this process of model/data comparison takes place at different levels. On the other hand, insofar as KT can be seen as an algorithmic rephrasing of IIT, it maintains some of IIT's relationships with anatomy and physiology.

2.4.5 Implication for empirical measures of consciousness

The first implication of KT is that we should study data and frame experimental questions using algorithmic complexity inspired metrics. An existing example is that of LZW compression (Cover and Thomas, 1991; Ziv and Lempel, 1978) of EEG data, already used with success (Casali et al., 2013; Schartner et al., 2015; Zhang et al., 2001; Zhao et al., 2007). Other estimators of K may provide better metrics in some scenarios.

As pointed out, the algorithmic complexity of inputs in evoked potential studies, of brain data (compressibility across space and time, spontaneous or perturbed), behavior, and the algorithmic mutual information between data and behavior should all provide relevant measurements in this framework.

In addition, there are probably fundamental links between algorithmic aspects of a system and derived quantities such as scale-free behavior or integrated information (phi), which need to be studied.

In this sense, KT may provide a meta-framework in which to describe different metrics (LZW, entropy, power laws, fractal dimension, avalanches) and empirical venues to study consciousness, perhaps in combination with machine learning from such features.

3 Discussion about models of consciousness

3.1 Functional integration and differentiation in thalamocortical networks

There are both similarities and differences between the frameworks described above and a systematic comparisons between them would entail an analysis (including a philosophical argumentation) that is clearly beyond the scope of this document. In the context of Luminous and its practical scopes, it may be useful to sketch a scheme whereby the different models are classified and compared based on a limited number of items, or axes (*Table 3-1*).

A fundamental distinction is whether a model emphasizes the importance of integrative processes, the degree of differentiation (or the amount of information) in neural activity, or the, balanced conjoint presence of both neural integration and differentiation. This fundamental choice affects, in turn, the kind of measures that the different models propose, as well as their preference for specific anatomical structures.

In this respect, the GNW theory seems to focus primarily on integration, as often indicated by the term 'ignition'. This can be operationalized as a sudden and global activation of workspace neurons concentrated in fronto-parietal circuits that broadcast the message globally. Although the term information may appear in the GNW formulation, the underlying assumption is that the information is *in the message* being broadcast, and that it becomes conscious depending on how many neurons can access it. In this perspective, the main process underlying consciousness remains the 'ignition' of frontal-parietal networks (as assessed by the presence of a P3b or large-scale synchrony) but no explicit measures of information are formulated.

Instead, the KT seems to shift significantly along the integration-information axis, clearly focusing towards the latter. In KT, information refers to the messages/signals traveling in and out of the brain or even within parts of the brain (I/O streams), and to Shannon's definition of the information conveyed by those messages. In particular, KT suggests that the algorithmic mutual information between brain chatter and input streams will be high in a healthy, conscious brain. Thus, unlike GNW, KT proposes information-based measures but it does not explicitly index the extent of integrative processes (such as the presence/strength of large-scale interactions within the system). Regarding the possible anatomical substrate of consciousness KT seems to privilege the temporo-parietal-occipital junction, an association area that receives multimodal information.

Unlike the GNW and the KT theories, the DCH focuses explicitly on the joint presence of information and integration. The neural substrate of consciousness is thus defined as large functional cluster of neuronal groups that together constitutes, on a time scale of hundreds of milliseconds, a unified neural process of high complexity termed the "dynamic core" in order to emphasize both its integration and its constantly changing activity patterns. Accordingly, the DCH proposes a metric (C_N) that captures, at once, the presence of strong interactions and the differentiation of spontaneous neuronal

activity. Regarding the possible anatomical substrate of consciousness the DCH seems to focus on coalition of neurons that can involve the entire thalamocortical system with no preference for particular structures.

The IIT builds up on the DCH and differs from the GNW and the KT in a similar way, as it rests on the fundamental postulate that the physical substrate of consciousness is a system with a high capacity for integrated information. Compared to DCH, IIT further specifies this notion by drawing a tight link between information and causation. Indeed, by the IIT perspective, in order to index a system's capacity for consciousness, one should measure the amount of intrinsic information that the system can generate through internal causal interactions. Thus, the measurements proposed by IIT tend to be based more on the responses of the system to direct perturbations (of corticothalamic circuits) rather than on the observation of its ongoing dynamics, or on sensory evoked responses. Regarding the possible anatomical substrate of consciousness, the IIT focuses on the thalamocortical system with a preference for a so-called "posterior hot zone" in the parieto-occipital cortex.

	WHAT	WHERE	HOW	EMPIRICAL PROXY
DCH	*information & integration*	whole **Thalamo-Cortical** system (moving/shifting)	analysis of spontaneous activity	C_N, C_d
GNW	*integration*	**Fronto-Parietal** network	analysis of responses to *sensory* stimulation and of spontaneous activity	P3b, wSMI, γ-synchrony
IIT	*information & integration*	whole **Thalamo-Cortical** system (posterior cortex)	analysis of responses to *direct* perturbations	PCI, bistability
KT	*information*	**Temporo-Parietal** junction	analysis of responses to *sensory/tDCS* stimulation and of spontaneous activity	Kolmogorov complexity

Table 3-1. Comparison among the considered models of consciousness on the following items: what do they measure, which brain networks are mainly concerned (where), which data type they rely on (how) and which experimental measure better approximates each model (empirical proxy). DCH = dynamic core hypothesis; GNW = global neuronal workspace; IIT = integrated information theory; KT = Kolmogorov theory

3.2 From theories to computational modeling

In this section, we briefly summarize some of the key elements of the above-described theories of consciousness and outline how these elements can be taken into account in the design of a neurophysiology-inspired computational model aimed at improving the interpretation of human brain data recorded under various states of consciousness.

As stated above, three theories (DCH, GNW, IIT) claim that the thalamocortical circuits play a crucial role in the maintenance of consciousness or in the transition from one state to another. They also emphasize the concepts of wakefulness and awareness as two major dimensions of consciousness. In addition, other important aspects (like distributed neural activity, reentrant interactions, connectivity patterns, surrounding/feed-forward/feedback inhibition, dynamic reorganization of functional networks, rhythms and oscillations, couplings, frequency bands) are differently addressed in these theories. All these elements provide crucial guidelines for developing a computational brain model in the context of the Luminous project.

3.2.1 Wakefulness

3.2.1.1 Which type of electrophysiological activity could be modeled?

The computational model should be able to simulate different electrophysiological markers at the cortical level such as Alpha/Beta rhythm (wake stage W), theta rhythm (N1), Spindle and K complexes (N2), and Delta waves (N3). The shift between the different stages of sleep should be dependent upon the degree of inhibition of TC (thalamo-cortical) cells, and therefore upon the level of RTN interneurons firing rate controlled by brainstem activity. A sensory input should also be able to induce a switch from deep thalamic inhibition to an awake state.

3.2.1.2 Which neurophysiological circuits could be simulated?

Thalamus. The model will include the reticular thalamic nucleus (RTN) composed of a population of GABAergic interneurons targeting TC cells. These reticular interneurons will receive inputs from the cortex, from TC cells and brainstem. The TC cells will receive input from brainstem, RTN and cortex and project onto the cortical layer 4. The thalamo-thalamic projection should be able to produce an up-and-down state, depending on the degree of inhibition from the RTN (and therefore the interneurons firing rate). The higher the activity of inhibitory cells, the higher the up-and-down activity of TC cells which will mimic different level of wakefulness as encountered in distinct sleep stages.

Thalamo-Cortical projection. Thalamocortical Feedforward inhibition (FFI) is one of the most fundamental organizational elements in the brain. It implements at its core a temporally structured normalization process of input to a circuit. Empirical and modeling work suggests that, within a generic FFI motif, a normalizing function selects synchronous inputs and facilitates their propagation. One of the functions of FFI motifs is to regulate circuits information transfer between the thalamus and the cortical layer 4, but also to modulate cortico-cortical inter-area interactions (Akam and Kullmann, 2014;

Lee et al., 2014; Womelsdorf et al., 2014). The key components of the dynamic FFI motif are (1) strong excitatory connections to inhibitory neurons with less excitation to principal cells from the same source, and (2) an inhibitory forward connection to principal cells with a time constant that imposes a temporal structure on the FFI motif. This type of inhibition is also involved in the generation of the cortical gamma rhythm. Parvalbumine positive somatic-projecting GABAergic interneurons (PV+) participate indeed to cortical gamma oscillations generation through 1) thalamocortical feedforward inhibition in layer 4, 2) feedback inhibition in layer 2/3, and 3) *via* direct PV-PV coupling through electrical gap-junctions (Womelsdorf et al., 2014). This scheme efficiently extracts population-coded information among concomitant asynchronous inputs in situations when this input is oscillatory. Therefore, FFI represent a key element of consciousness regarding its role in gamma oscillation generation, inter-area communication, sensory integration and cortical inhibition. Each cortical population contains excitatory glutamatergic pyramidal cells, and at least two types of GABAergic interneurons (e.g. somatic-projecting interneuron, GABA fast and dendritic projecting interneurons, GABA slow) TC cells project onto several populations of cortical neurons. Glutamatergic projection coming from TC cells target both cortical pyramidal cells (feedforward excitation) and perisomatic-targeting PV+ interneurons (feedforward inhibition).

3.2.2 Awareness

Awareness is likely to be dependent on corticocortical functional connectivity. It is well established that synchronization affects communication between cortical neuronal groups. Each activated neuronal group (belonging to the same module) has the property to oscillate. Functional connectivity requires resonant oscillations between modules of the network. Rhythmic synchronization creates sequences of excitation and inhibition that focus both spike output and sensitivity to synaptic input to short temporal windows (Fries, 2015). Anatomical connections are dynamically rendered effective or ineffective through the presence or absence of rhythmic synchronization, in particular in the gamma band (30–90 Hz).

3.2.2.1 Which type of electrophysiological activity could be modeled?

No functional connectivity during the up and down state ("sleep mode"). When the thalamocortical input is a slow, synchronized burst discharge (as during the up and down state), strong feedforward inhibition combined to lateral inhibition between cortical populations should provide a spot of excited pyramidal cell surrounded by inhibited pyramidal cells (surrounding inhibition), a decrease of inter-area communication and therefore a decrease of functional connectivity (decrease of awareness).

Gamma oscillations and Inter-area functional connectivity ("awake mode"). At the cortical level, several neural populations should be able to oscillate in the gamma range (30–90 Hz). Several distinct cortical neuronal populations have to be interconnected to each other. Connectivity between different populations of cortical neurons should allow phase-locking of gamma oscillations and emerge when the thalamocortical loop is tuned in the "awake" mode (e.g. phasic/tonic discharge).

3.2.2.2 Which neurophysiological circuits could be simulated ?

The thalamocortical and the corticothalamic connections of each population should be present in the model and considered between each cortical population and the thalamic cells.

Gamma oscillations In order to produce gamma oscillations, at least three main type of neurophysiological connections should be present in the model (Womelsdorf et al., 2014):

- mutual inhibition of somatic-projecting GABAergic interneurons (fast-spiking interneurones) through electrical gap-junctions,
- feedback inhibition from somatic-projecting GABAergic interneurons, and
- thalamocortical feedforward inhibition through thalamic activation of cortical somatic-projecting GABAergic interneurons.

Functional connectivity. Several distinct cortical neuronal populations have to be interconnected between each other through excitatory, long-range pyramidal cells projections onto other targeted pyramidal cells. The connectivity pattern (strength, convergence, divergence) should be tuned according to the neurophysiological data available in the literature. Pyramidal cells of one neuronal population should also project onto distant somatic-targeting GABAegic interneurons to account for 1) the possibility to trigger gamma oscillations by activation of mutual inhibition, and 2) corticocortical lateral inhibition. Based on this connectivity pattern, generation of gamma oscillations in some of these neuronal populations should create rhythmic sequences of excitation and inhibition that should open a temporal windows to bind activity of distant networks.

Vertical and horizontal circuits. Ideally, the model should be able to promote different sleep-wake stages (depending on thalamic output). In the sleep mode, large and slow discharges of neuronal cortical populations surrounded by a large inhibitory barrage should induce specific biomarkers on the simulated EEG such as slow waves. In this case, corticocortical functional connectivity is expected to be low. In the awake mode, distinct types of TC cells discharge should be associated with normal background frequency band in the simulated EEG and the spontaneous emergence of gamma oscillations. The decrease of feedforward inhibition (FFI) and lateral inhibition might favor functional connectivity between distinct neuronal groups, and therefore increase the level of awareness.

4 Metrics of consciousness

So far, different quantitative measures of consciousness have been proposed (*Figure 4-1*), each related to a specific model. All these measures can be roughly classified according to what they mainly address, being either information, or integration, or both.

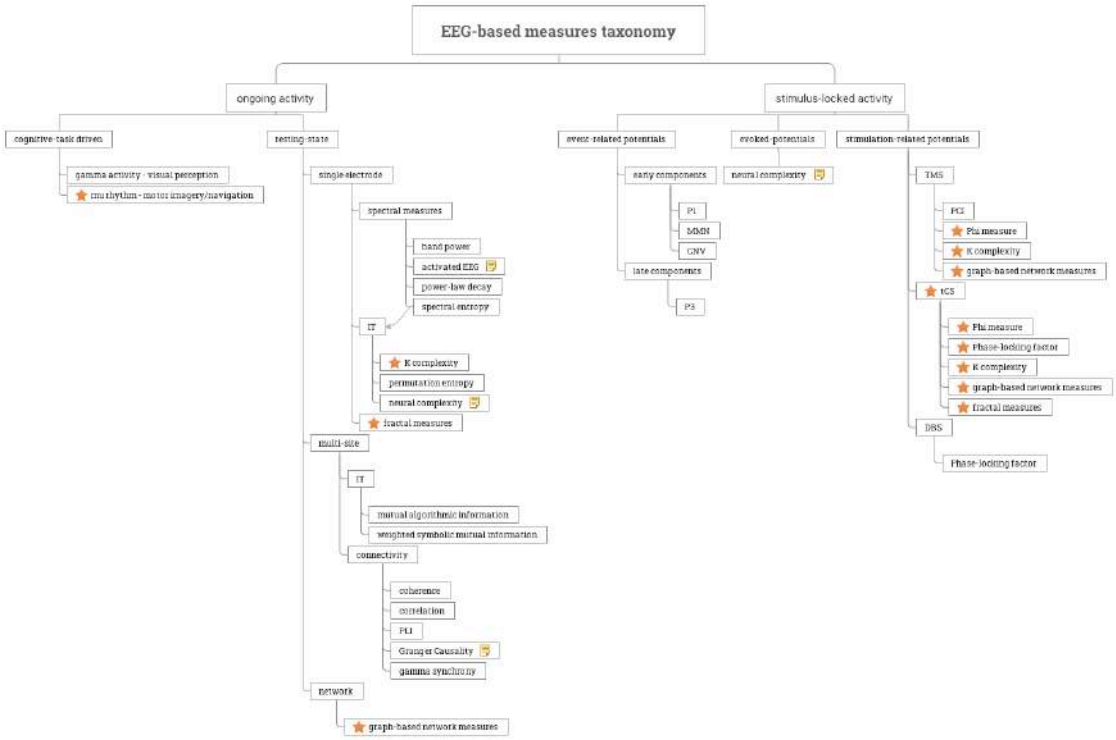

Figure 4-1. Taxonomy of different EEG-based univariate consciousness metrics discussed in this deliverable based on the experimental methodology used for the EEG acquisition. Star-labeled metrics have not been explored in the currently existing literature and are planned to be used in Luminous works.

4.1 Information-based metrics

4.1.1 Activated EEG

Activated or desynchronized EEG, consisting in low-voltage fast activity was first considered a candidate index of consciousness because it is typically observed during attentive wakefulness (Moruzzi and Magoun, 1949). The mechanisms behind the transition from the low-voltage fast activity during wakefulness to the high-voltage, slow activity during deep slow wave sleep anesthesia have been extensively investigated (Steriade, 2000). When thalamic neurons are hyperpolarized, they switch from a tonic to a bursting firing mode, resulting in the synchronization of the EEG in the spindle (12–14 Hz) and theta (5–8 Hz) ranges (Steriade, 2000). Slow waves, i.e. large oscillations in the delta range <4 Hz, are generated when cortical neurons start

alternating synchronously between depolarized up‐states and hyperpolarized down-states (Steriade et al., 2001). Loss of consciousness is associated to this kind of activity pattern in several physiological, pharmacological and pathological conditions (Brown et al., 2010). Occurrence of slow waves may be due to a decrease in the firing rate of subcortical activating systems (McCormick et al., 1993; Moruzzi and Magoun, 1949), by excessive inhibition exerted by the globus pallidus on the thalamus (Schiff, 2009) or by a critical level of cortical deafferentation (Timofeev et al., 2000) common in comatose patients (Fernández-Espejo et al., 2011).

Nevertheless, the use global EEG patterns does not always reliably distinguish between conscious and unconscious patients. Moreover, it is true that electrical stimulation of the midbrain reticular formation elicits EEG activation and behavioral awakening in a cat (*Figure 4-2A*) (Moruzzi and Magoun, 1949): low-voltage fast-activity replaces slow waves whenever cortical neurons cease to alternate between depolarized up states and hyperpolarized down states to attain a steady membrane potential (Steriade et al., 1996). However, intracranial electrical recordings have shown that unconscious patients during NREM sleep may show local EEG activation in the sensorimotor cortex (*Figure 4-2B*) (Nobili et al., 2012). On the other hand, rhythmic bilateral slow waves may be present in awake, conscious patients in some cases of non-convulsive status epilepticus (Gökyiğit and Calişkan, 1995; Vuilleumier et al., 2000) (*Figure 4-2C*). The presence of higher EEG frequencies *per se* can be misleading, since widespread, low-amplitude alpha rhythms may be recorded in patients with in a severe post-anoxic coma (*Figure 4-2D*) (Brenner, 2005).

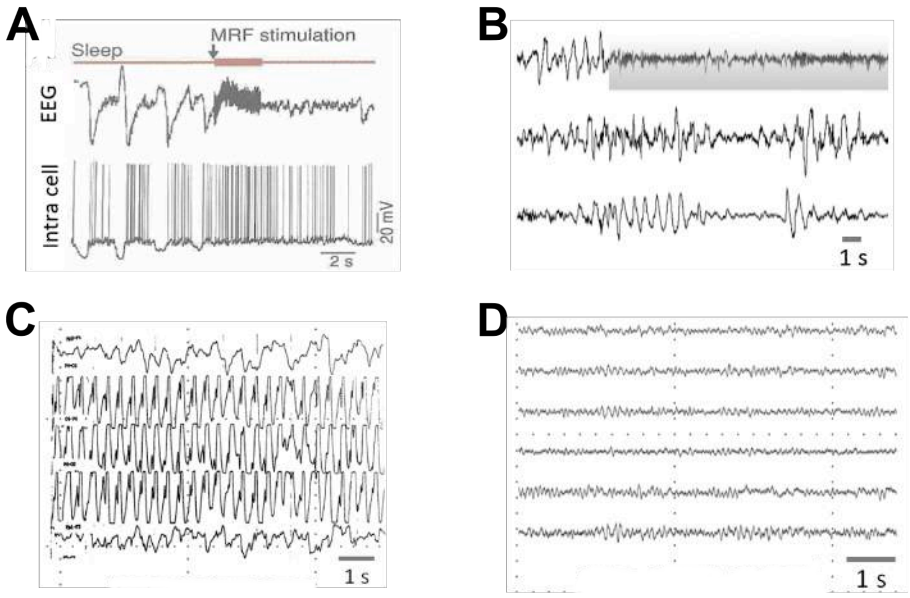

Figure 4-2. Results obtained with activated EEG in different conditions: (A) stimulation of the midbrain reticular formation in cats, (B) unconscious subjects during NREM (non-rapid-eye-movement) sleep, (C) non-convulsive status epilepticus, (D) post-anoxic alpha coma.

Potential relevance for Luminous. *The evaluation of activated EEG does not require advanced technical and computational resources, therefore could be easily applied in different experimental settings. Previous studies have shown that looking at activated*

EEG alone is not enough to characterize different states of consciousness. Nonetheless, activated EEG suggests that a large repertoire of neural activity patterns (differentiation) is important for consciousness and therefore could be reconsidered in a multi-modal assessment for obtaining complementary information about the overall state of brain circuits.

4.1.2 Kolmogorov complexity metrics (K)

Algorithmic complexity provides a horizontal "technology" for Luminous. It provides the framework for computation of algorithmic complexity measures such as LZW and methods such as PCI. It can be employed to characterize the complexity of sensorial inputs and behavioral outputs in experiments, or to measure the "integration" of, e.g., all the electrode data in a dataset. Here we describe the main concepts and approximations to measurement in practice, as well as future lines of work for Luminous.

Compression and therefore simplicity were formalized by the notion of algorithmic complexity or Kolmogorov complexity (K), a mathematical concept co-discovered during the second half of the 20th century by Solomonoff, Kolmogorov and Chaitin[1] which provides a formal cornerstone to address the question of compression\. We recall its definition: *the Kolmogorov or algorithmic complexity of a data set is the length of the shortest program capable of generating it.*

More precisely, let U be a universal computer (a Turing machine), and let p be a program (Cover and Thomas, 1991; Li et al., 2008). Then the Kolmogorov or algorithmic complexity of a string x with respect to U is defined by

$$K_U(x) = \min_{p:\, U(p)=x} l(p),$$

the minimum length over all programs that print the string x and then halt. An important fact is that this is a meaningful definition: although the precise length of the minimizing program depends on the programming language used, it does so only up to a constant. That is, if U is a universal computer, then for any computer A and all strings x we can easily show that $K_U(x) < K_A(x) + c$.

Gödel's incompleteness theorem, or its equivalent, Turing's halting theorem, implies we cannot compute in general K for an arbitrary string: it is impossible to test all possible algorithms smaller than the size of the string to compress, since we have no assurance that the Turing machine will ever halt (Chaitin, 1993). However, if we define a compression scheme and allow for a finite computation time, compressibility becomes a practical question. A theory of 'practical' complexity can tell us what the expected length of the shortest programs can be if computational resources are finite (in time and space), or if there are constraints in natural compressing mechanisms. An excellent example of this is in the implementation of LZW (Cover and Thomas, 1991; Ziv and Lempel, 1978), used for compression of data files (in UNIX compress or in GIF files). LZW provides the starting point for exploring compression metrics in brain data. It is a simple yet fast algorithm that seeks to exploit the recurrence of patterns in data streams.

[1] Kolmogorov complexity is also known as 'algorithmic information', 'algorithmic entropy', 'Kolmogorov- Chaitin complexity', 'descriptional complexity', 'shortest program length' and 'algorithmic randomness'.

It is limited in the sense that it cannot compress random-looking data generated by simple programs (e.g., the sequence binary digits of π).

A further connection between the notion of simplicity and K was developed by Solomonoff (Li et al., 2008) with an emphasis on statistics and prediction. The fundamental quantity here is the algorithmic or universal (un-normalized) probability PU (x) of a string x. This is the probability that a given string x could be generated by a random program. An important result is that this is given by $PU(x) \approx 2^{-K_U(x)}$. Thus, short programs contribute most to the probability of observing a given data string and, furthermore, the probability of a given string to be produced by a random program is dominated by its Kolmogorov complexity. Summarizing, the probability of observing a string is actually dominated by the length of shortest program capable of generating it. Simplicity is therefore a good strategy for prediction if the universe is generated by random program generating mechanisms. The precept that short explanations are in some sense more likely is the essence of the Minimum Description Length (MDL) and the Minimum Message Length (MML) approaches to statistical inference (Li et al., 2008; Vitanyi and Li, 2000; Wallace and Dowe, 1999). Among possible explanations for data (an observed data string), those which are shortest (program + error or "remainder" data) are more likely.

At this point it is also important to mention the concept of Universal (Levin) Search (US), which resolves the problem of computability of K and adds an element of practical relevance by penalizing slow programs (Hutter, 2007). This modified complexity measure is actually computable, and essentially equivalent to the logarithm of the running time to produce the desired string. Conceptually, it may be a relevant element in computation in life, where time for computation and memory requirements will play a factor driven by natural selection. A related concept is Logical Depth (Bennett, 1988; Zenil et al., 2016), where the complexity of a string is defined by the time that an computation process takes to reproduce the string from its shortest description. However, these definitions build on Kolmogorov complexity rather than supersede it in a fundamental way.

An important derived concept we will make use of is the mutual algorithmic (Kolmogorov) complexity between two strings (Domingos, 1999). We recall first the definition of mutual (Shannon) information (Ludwig and Falter, 2012) between random variables X and Y , $I(X; Y) = H(Y) - H(Y|X)$, where H(X) is the entropy of X ($H(X) = -E[\log P(X)]$) and H(Y|X) is the entropy of Y conditioned on X ($H(Y|X) = -E_{p(X,Y)}[\log P(Y|X)] = H(Y,X) - H(X)$). Similarly, the mutual algorithmic information (MAI) between two strings x and y is given by (Grunwald and Vitanyi, 2004; Li and Vitanyi, 1997) $IK(x : y) = K(y) - K(y|x)$. To approximate this quantity using real data we can use the fact that $K(x,y) \approx K(x) + K(y|x)$ so that $K(y|x) \approx K(x,y) - K(x)$ and $IK(x : y) \approx K(y) + K(x) - K(x, y)$. Each of the terms in the right hand side can be estimated using LZW, for example (as discussed below).

Finally, there exist links between algorithmic complexity and other notions of complexity from the theory of non-linear dynamics and chaos that require further research (Prokopenko et al., 2009) and which we will discuss further below.

Approximations of K exist, including the aforementioned LZW compression, used in both spontaneous and perturbed brain consciousness metrics. However, LZW is a limited approximation, and will miss certain types of regularities. We discuss next proxies for K and mutual K beyond LZW (see Ruffini 2016, *submitted*).

Machine learning approaches using neural networks: Neural networks have already been used for image compression by using as auto-encoders (Jiang, 1999). Similarly, EEG data from a given cohort (e.g., MCS or UWS) can be supplied to a NN to train it for sparse auto-encoding. The sum of the state data at the "thinnest" layer plus the error (which can itself be then compressed by LZW) provides a new estimate of K. Presumably, we will find different compressibility ratios in different scenarios. Similarly, we can also employ RNNs (Echo state networks (Jaeger, 2001) or LSMs) for prediction (as in sequence entropy) to compress temporal data.

Genetic programming for compression: in a more exploratory fashion, in (Ruffini, 2007) an approach using CAs was proposed to find short programs (since CAs such as Rule 110 are universal computers) and thus estimate K. We can extend those ideas to search for short program strings and compress program plus error using LZW. More generally, other forms of genetic programming provide yet another method to consider to estimate K.

Potential relevance for Luminous: *The core concept in KT is algorithmic information. For this reason it is important to use and develop tools to estimate the algorithmic complexity of data in all experimental scenarios. LZW remains a basic tool for this, but other metrics using machine learning should be developed and tested in the project. K metrics can be used to quantify complexity of input data (e.g., sensory data), brain data, behavior, and mutual algorithm information in various dimensions.*

4.1.3 Scale-free approaches: Consciousness and criticality

Self-organized criticality (SOC) is claimed as a fundamental property for the functioning of the brain where the latter actively retunes itself to critical states through active decentralized processes allowing for optimal information coding (Hesse and Gross, 2014). Hence, several researches used SOC hallmarks to explain and characterize different consciousness states. An intrinsic feature of criticality is the scale-free dynamics detected by the presence of power-law behaviors in neuronal data.

4.1.3.1 Neural avalanches

To estimate whether the neuronal dynamics in the human brain operate close to criticality, or in sub- or supercritical states, spatio-temporal clusters of enhanced activity called neuronal avalanches are extracted from neuronal signals. The SOC state is detected when the size distribution of these neuronal avalanches follows a power-law (scale-free neuronal avalanches) (Beggs and Plenz, 2003). The size s of a neuronal avalanche is defined as the total number of recording sites that show enhanced activity during one avalanche. In this context, an enhanced activity is detected in a given potential when the area under the positive deflection lobes between two zero crossings exceeds a preset threshold (Beggs and Plenz, 2003).

In (Priesemann et al., 2013), neural avalanches size was analyzed across different vigilance states in humans, using local field potentials recorded with intracranial depth electrodes from epileptic patients. It was shown that avalanche distributions differed slightly but consistently between vigilance states. These differences were due to

deviations from power-law scaling rather than changes in critical exponents. In fact, by fitting avalanche distributions to a function of the form $F(s) \sim e^{-\alpha s} s^{-\tau}$, slow-wave sleep (SWS) showed the largest avalanches measures ($\alpha = 0.0064$, $\tau = 1.52$, close to criticality state); whereas wakefulness showed intermediate ones ($\alpha = 0.0415$, $\tau = 1.32$) and rapid eye movement sleep (REM) showed the smallest ($\alpha = 0.0573$, $\tau = 1.24$, subcritical states). This conclusion was confirmed by Meisel et al. in (Meisel et al., 2013) where they established that signatures of criticality are progressively disturbed during wake and restored by sleep. They demonstrated that the precise power-laws governing the cascading activity of neuronal avalanches and the distribution of phase-lock intervals in human EEG recordings are increasingly disarranged during sustained wakefulness. Conversely, sleep restores the critical state resulting in recovered power-law characteristics in activity and variability of synchronization.

The above cited studies about criticality in the brain are focused on the spatial or structural complexity. Yet, there is also the temporal complexity where the focus is on the time long-range correlations with power-law decay (equivalent to 1/f noise) and on time intermittency.

4.1.3.2 Power-law decay (1/f noise)

In the field of EEG, most attention is generally dedicated to study rhythmic activities of different frequency bands. Recently some studies underlined the arrhythmic behavior of EEG signals, stimulated by the concept of scaling laws. Indeed, EEG spectral power $P(f)$ estimates may obey scaling laws of the form $P(f) \approx f^{\beta}$. Thus taking the log of both sides of the expression yields a straight line with slope β. A typical example is given by Kello et al (Kello et al., 2010): fluctuations of the amplitude envelope of MEG in an extended alpha-band (6.7 – 13.3 Hz) follow a $1/f$ relation at least for a wide frequency band down to infra-slow frequencies.

The estimation of the power-law exponent of a given EEG spectrum is not trivial. To realize this, He et al (He et al., 2010) used the method of Yamamoto and Hughson (Yamamoto and Hughson, 1991), that permits separating the oscillatory frequency components from the scale-free components: the so-called coarse-graining spectral analysis. An example, showing power spectra of intracerebral EEG signals of three patients in two conditions, awake and SWS, is shown in *Figure 4-3*.

Generally, *1/f* spectrum is interpreted in time domain as a *scale-free* property where a signal $X(t)$, said to be *self-similar*, shares the same statistical properties with its own dilated and scaled version $a^{-H}X(at)$. Here, H is the self-similarity parameter referred also as *Hurst exponent*. Time series with *scale-free* behavior or "fractal dynamics" (He et al., 2010) show a long-range autocorrelation function that decays slowly as function of time lag. In a more general way, the q^{th} statistical moment of such process follows a power-law distribution of the form

$$\mathbf{E}|X(t)|^q \propto |t|^{qH}.$$

Several methods were proposed to estimate the *Hurst exponent*; however, the most popular are the Detrended Fluctuation Analysis (DFA) and the Wavelet Leader Based Multifractal Formalism (WLBMF). The former method estimates the exponent from the 2nd order moment after removing the linear trend from the signal using a least-square fit in order to make the analysis less sensitive to false correlation induced by trends persisting over longer time-scales (Ihlen, 2012). The latter uses wavelet leader

formalism to estimate *Hurst exponent* and is more robust to non-stationarity issues (Wendt et al., 2007). Free MATLAB toolboxes exist for both methods and will be tested in Luminous project.

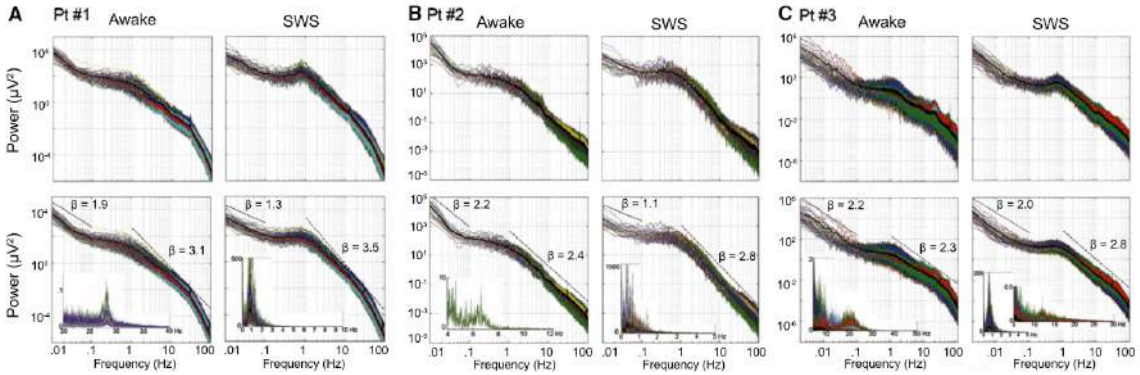

Figure 4-3. Power spectra in EEG data. This figure illustrates the fact that the spectra present rhythmic components (e.g. slow oscillation, sleep spindles, theta, alpha, beta, and gamma frequency components), superimposed on an arrhythmic component that follows approximately the $1/f^{\beta}$ power-law, but with different values of the exponent β depending on frequency range, with a "shoulder" between the low and high ranges. [Adapted from (He et al., 2010)]

In some cases where the analyzed data are highly nonstationary or non-Gaussian, scale-free temporal dynamics are no longer better explained by self-similarity and multifractality is rather used. The latter enables to account for local fluctuations over time around the *1/f* slope. Hence, instead of using *Hurst exponent H* (a global parameter), a local power-law exponent *h* is used (*Hölder exponent*). The normalized probability distribution of *h* in log-coordinates is defined as the *multifractal spectrum D* (Lopes and Betrouni, 2009). This multifractality amount can be measured using box-counting (Accardo et al., 1997), Multifractal DFA (MFDFA) and WLBMF. It is noteworthy that, in the case of self-similarity, the *multifractal spectrum D* is reduced to a single *Hölder exponent h* that coincides with *Hurst exponent H.*

Several works have reported the modulation of self-similarity and multifractality attributes between contrasted conscious states, namely sleep vs. awake. In (Weiss et al., 2009), a spatiotemporal analysis of self-similar and multifractal properties of whole-night sleep EEG recordings was carried out. *D* values showed a negative, while *H* values exhibited a positive correlation with sleep depth. This indicates that EEG signals tend to be less multifractal and have longer memory properties during NREM4 compared to NREM2 and REM sleep stages.

In (Leistedt et al., 2007), the objective of the study was to find out whether sleep EEG abnormalities are state-dependent or scar-marker in patients with major depressive disorder (MDD) in full to partial remission. *Hurst exponent H* was computed using DFA and during the different sleep stages. It has been noticed that, in both healthy and MDD groups, *H* values increased from light sleep (0.80–0.96) to deep sleep (0.83–1.09) and decreased during REM sleep (0.66–0.93). Here, values are expressed as minimum-maximum range. These results were also confirmed in (Lee et al., 2004) where awake and REM presented the same *H* values (1.08 + 0.17) while N1, N2 and N3-4 (SWS) showed increased values (1.19 ± 0.22), (1.20 ± 0.20) and (1.31 ± 0.12) respectively. All the differences were significant at the 0.05 level of significance.

In (Tagliazucchi et al., 2013), the temporal memory of BOLD signals across the human sleep cycle (wakefulness and N1, N2, and N3 sleep) was studied. Authors hypothesized that a breakdown of the long-range temporal correlations occurs during the descent to deep sleep. Changes in the voxel-wise spatial distribution of H of EEG-fMRI BOLD signals across all stages of the human NREM sleep cycle were analyzed using DFA. The effect of sleep stage on H was highly significant (p-value $< 10^{-9}$). Post hoc tests revealed significant differences between both wakefulness and all other sleep stages (p-value $< 10^{-3}$ in all cases). A widespread decrease of H was observed for wakefulness vs. N2 and N3 sleep parietal and frontal regions associated with the default mode network (a set of task-deactivated regions implied with internal conscious cognitive processes) and attention resting state networks. Additional decreases were located in the inferior temporal cortex and thalamus.

The modulation of self-similarity and multifractality parameters was also explored in the case of perceptual learning framework (Buiatti et al., 2007; Zilber et al., 2012, 2013). Using MEG signals acquired from 24 healthy subjects undergoing a perceptual task (visual and audiovisual), Zilber et al. have measured H and D (using WLBMF) in some regions of interest, and this before and after training during task and at rest. Results showed a reduction of self-similarity in evoked activities interpreted as an increase of the neural excitability which allows the participants to respond more quickly after the stimulus onset.

In spite of what was formerly reported , other researches claim that $1/f^{\beta}$ spectra are not necessarily associated with critical states (De Los Rios and Zhang, 1999). Most relevant in the present context is the study of Bédart et al (Bédard et al., 2006), who investigated whether $1/f^{\beta}$ power scaling can be demonstrated in Local Field Potentials (LFPs) *in vivo* (in cat). During wakefulness the power spectra shows two scaling regions ($1/f$ between 1 and 20Hz, and $1/f^3$ between 20 and 65Hz); during SWS the $1/f$ at the low frequency band is masked, but the $1/f^3$ is still present within the higher frequency band. The conclusion is that $1/f$ scaling reported for some EEG signals is also present in LFPs during waking but only for some specific frequency bands. It is especially noteworthy that these authors also analyzed time series of action potentials of cortical neuronal activity (Inter-Spike Intervals). The latter showed exponential distributions typical of Poisson processes, during both waking and SWS (although with different slopes). Further these authors investigated whether signs of criticality could be identified using the same method as reported by Beggs and Plenz (Beggs and Plenz, 2003); they found, however, that the distributions of avalanches in these spike series did not follow power-law scaling, rather they fitted exponential distributions. Most interesting they conclude by showing using a simple biophysical model that in order to obtain a $1/f$ scaling of LFPs, it is sufficient to take into account the biophysical fact that the extracellular currents underlying the LFPs have to cross the extracellular space, which has both resistive and capacitive properties. This constitutes a resistor-capacitor filter that has, on its own, a classic $1/f$ scaling property; this property confers the $1/f$ spectral characteristic to the LFPs. These conclusions do not support the hypothesis that $1/f$ scaling of EEG signals power spectra is evidence for the existence of neuronal critical states, rather they point out the importance of taking into account straightforward biophysical properties in interpreting LFPs and EEG signals.

4.1.3.3 Time intermittency

Time intermittency is defined by the presence of crucial events in the brain. Crucial events here denote the abrupt transitions or Rapid Transition Processes (RTPs) from and to metastable states, via multichannel EEGs. RTPs were exploited to investigate time intermittency in the brain. It was established that time intervals τ between two consecutive crucial events or called also Waiting-Times (WT) follow a power-law distribution $F(\tau) \sim 1/\tau^{\mu}$, where μ denotes the intermittency exponent (Allegrini et al., 2009). Correlation between this fractal intermittency and consciousness stream was investigated in (Lee et al., 2002; Paradisi et al., 2013) where Paradisi et al. had estimated the intermittency exponent μ for different sleep phases. The estimation of μ was indirectly evaluated through event-driven diffusion scaling of EEGs where the second moment $\sigma(t)$ of the diffusion process exhibits a power-law behavior of scaling $H(\mu)$, a function of the intermittency exponent μ. The estimation of $H(\mu)$ using DFA gives access to the values of μ (analytical expressions of H as a function of μ were determined in the case of WTs with power-law distribution). The diffusion scaling H (consequently μ) showed different values according to the wakefulness state (see *Figure 4-4*). During pre-sleep (wake) and REM phases, H was evaluated to 0.75 ($\mu = 2.5$), whereas in deep sleep (SWS), it was estimated to 0.5 ($\mu > 3$). H and μ could be then proposed as reliable correlates of consciousness.

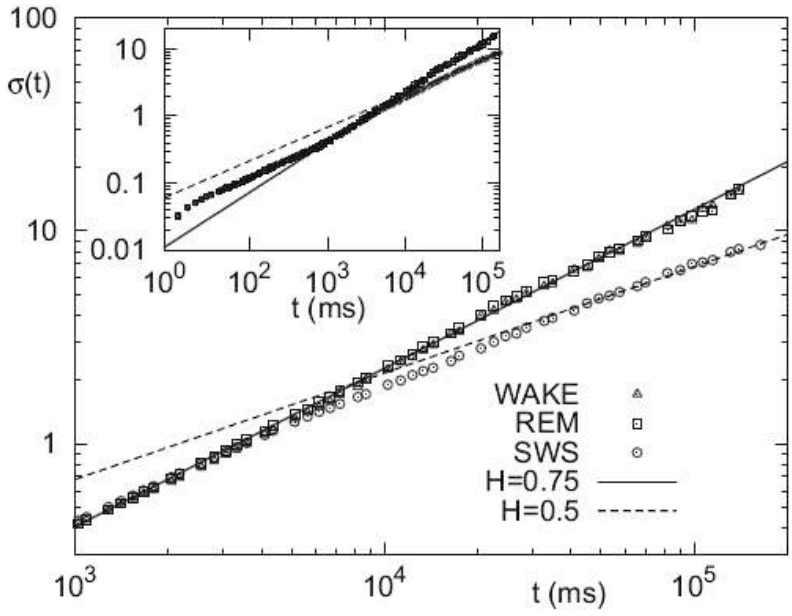

Figure 4-4. Asymptotic time range in the DFA applied to different sleep phases (cycle I). Continuous and dashed lines are a guide to the eye for the slopes H = *0.75 and* H = *0.5, respectively. In the inset, the entire time range over which the DFA has been computed is reported. [Adapted from (Paradisi, Allegrini et al. 2013)]*

Potential relevance for Luminous. *Scale-free metrics can be used in all the experiments involving EEG measures, especially where the sleep stages are studied (experiments conducted by ifADO).*

4.1.3.4 Fractal measures

Various processes in natural sciences, seem to be fractal or multifractal, which has many implications on their properties. In particular, a fractal set tends to fill the whole space in which it is embedded, and has a highly irregular structure while possesses a certain degree of self-similarity or self-affinity. The idea is that natural phenomena that seem to present a higher level of complexity, may look similarly complex under different resolutions, and although this structure may be initially seen as a complex one, it actually is a source of simplicity. This behaviour of a fractal can be captured by its fractal dimension, which can be regarded as a measure of complexity. Thus, a seemingly complex system can be explained by a relatively low set of parameters, such as its fractal dimension. Fundamental to most definitions of fractals is the idea of scale invariance, which refers to the fact that the relationship between the measurements and the scale must obey a power-law form, and assigns to them a scale-free behaviour. For instance, the temporal variations of EEG signals are typically non-stationary, and exhibit long-range correlations over many time scales, indicating the presence of self-invariant and self-similar structures (Lee et al., 2002).

Higuchi fractal dimension (Higuchi, 1988) is directly related to the fractal dimension of a signal, and estimates the length of a time series over a range of scales, which follows a power-law. The fractal dimension is then estimated from the slope of the line that fits the log-log curve between the length and the range of scales in a least-squares sense. Compared to other, more conventional, measures of estimating the fractal dimension, such as the correlation dimension, the advantages of Higuchi fractal dimension are that it is easy to implement and that the input time-series does not need to be embedded in a phase space. Higuchi fractal dimension has been used for the analysis of MEG signals from Alzheimer disease patients (Gómez et al., 2009).

Normalized length density (NLD) (Kalauzi et al., 2009) is another measure that estimates the complexity of a signal in very short epochs. The idea is to count the local extrema and normalize this number with respect to the total number of points, but doing this in a continuous way. Thus, the NLD estimates the length of a signal divided by the number of samples and normalized with respect to the average signal magnitude. It is a measure proportional to the local extrema density, and is related to the fractal dimension of a signal through a power law.

The dimension of the minimal cover estimates the local fractal characteristics of a signal (Dubovikov et al., 2004), thus represents a measure of its multifractality. Multifractals have some parts, which are reduced-size copies of the whole, thus they have different fractal dimensions in different parts, and the final fractal dimension is a measure of these local fractal characteristics. The dimension of the minimal covers introduces a local instantaneous parameter to follow the multifractality of non-stationary processes (such as EEG). The idea is that the signal is divided into equal intervals and the minimal covers are estimated as rectangles with length equal to each interval's length and height equal to the difference between the maximum and minimum values within this interval. Then the average of these differences for each scale follows a power law with respect to the scale, whose exponent is linearly related to the dimension of the minimal cover.

Potential relevance for Luminous. *Fractal measures are potentially important in the*

context of Luminous, as they capture the level of complexity of the input signals. Since the underlying hypothesis across various states of consciousness is that they affect the complexity of the brain, fractal measures are expected to provide important information on brain complexity, and to be able to discriminate across various stages of consciousness.

4.1.4 Permutation Entropy (PE)

Permutation entropy (PE) is a powerful tool for the analysis of time series where time information such as time causality and time scales are duly taken into consideration, giving access to the dynamic behavior of a given studied system such as periodicity, chaos or randomness. It has the quality of simplicity, very low computational cost and robustness to low signal to noise ratio compared to other similar methods (Bandt and Pompe, 2002).

PE is evaluated using the probability density function (PDF) of permutation patterns. The latter is built by dividing the temporal signal into sub-vectors of length L (the embedding dimension) of subsequent elements or of elements separated by τ samples (the embedding delay) which permits to map the signal dynamics at different temporal resolution. Each sub-vector is associated to an ordinal pattern defined as the permutation of the amplitude values sorted in an ascending order. The PDF of these permutation patterns are then estimated and the PE is deduced by applying Shannon's classical formula (Zanin et al., 2012).

PE has been prominently used to evaluate statistical complexity and detect spatio-temporal dynamic change in EEG signals. In (Jordan et al., 2008; Li et al., 2008; Olofsen et al., 2008), it was shown that PE can be used to efficiently discriminate between different levels of consciousness during anesthesia, providing an index of the anesthetic drug effect. In (Sitt et al., 2014), PE measured in the theta and alpha frequency ranges were proven an efficient parameter to discriminate vegetative state (VS) from the other groups (minimally conscious state MCS, conscious state CS and healthies). In general, a greater value of PE, especially over centro-posterior regions indicating a more complex and unpredictable distribution, indexed a higher state of consciousness.

Potential relevance for Luminous. *PE was proved an efficient metric when applied to EEGs of anesthetized patients and patients with disorders of consciousness (VS and MCS). Thus, it can be used in the experiments planned with UOXF, EKUT and ULG partners. It is nonetheless interesting and tempting to test it also in other different cases such as sleeping stages (ifADO).*

4.2 Integration-based metrics

4.2.1 Functional and effective connectivity of the cerebral cortex

As highlighted in section 2, brain connectivity is an essential and common characteristic in the models of consciousness proposed so far. In the context of the Luminous project, we aim at estimating brain connectivity from data recorded under various conditions corresponding to different states of consciousness.

In the assessment of brain connectivity using EEG/MEG recordings, an important aspect is the determination of the statistical relationships between signals, either in the "electrode space" (i.e. signals directly recorded at the level of scalp electrodes) or in the "source space" (i.e. signals reconstructed at the level of brain sources, mainly neocortical, from scalp EEG). Multivariate analytical methods provide the basic tools to estimate functional relationships of brain signals. As pointed out by Friston et al (Friston et al., 2013), functional connectivity measures address statistical dependencies between brain signals that are instantaneous or undirected. Approaches dedicated to assess effective connectivity, aim at providing information about the directed or causal influence of one system upon another one.

See the recent paper by (Bastos and Schoffelen, 2016) for a review of functional connectivity methods, including manners to address the problem of a) common referencing and b) common inputs to the data. We have also pointed out above mutual algorithmic information as an interesting new candidate to study connectivity.

4.2.1.1 A short overview of methods for assessing functional connectivity

The basic concepts on which statistical functional analysis are based, can be found in several comprehensive textbooks and reviews (Sanei and Chambers, 2008). In brief, functional connectivity can be estimated using linear and nonlinear analytical measures. Typically, linear methods like cross-power and phase spectra and the derived coherence functions have been used in a wide variety of research and clinical investigations of EEG signals for decades. In the case of scalp derivations, there are a number of hurdles that make the interpretation of these functions with respect to functional connectivity difficult or ambiguous, namely the reference electrode, the volume conductor effect and noise of different sources (EMG, EOG artifacts). Therefore whenever computing indices of functional connectivity it is recommended to use EEG bipolar derivations with closely spaced electrodes, without a common electrode, and it is even preferable to use Laplacian derivations or to compute these indices at the level of brain sources.

Beside classical coherence-based methods, other approaches developed over the past twenty years among which non-linear regression analysis, mutual information, phase synchronization (typically, the phase-locking value estimated using Hilbert phase entropy; Hilbert mean phase coherence; wavelet phase entropy or wavelet mean phase coherence) and generalized synchronization (estimated by various similarity indices or synchronization likelihood).

In the following, we provide the details of the most commonly-used measures of functional connectivity. The nonlinear correlation coefficient (h^2) is a non-parametric

measure of the nonlinear relationship between two time series *x* and *y*. In practice, the nonlinear relation between the two time series is approximated by a piecewise linear curve.

$$h_{xy}^2 = \max_{\tau}\left(1 - \frac{\mathrm{var}(y(t+\tau)/x(t))}{\mathrm{var}(y(t+\tau))}\right)$$

where $\mathrm{var}(y(t+\tau)/x(t)) \triangleq \arg\min_f\left(E[y(t+\tau) - f(x(t))]^2\right)$ and $f(x)$ is the linear piecewise approximation of the nonlinear regression curve.

The mutual information (*MI*) between signal *x* and *y* is defined as:

$$MI_{xy} = \sum p_{ij}^{xy} \log \frac{p_{ij}^{xy}}{p_i^x p_j^y}$$

where p_{ij}^{xy} is the joint probability of $x=x_i$ and $y=y_j$. In the case of no relationship between *x* and *y*, $p_{ij}^{xy} = p_i^x p_j^y$, so that the *MI* is zero for independent processes. Otherwise, MI_{xy} will be positive, attaining its maximal value for identical signals.

For two signals *x* and *y*, the phase locking value is defined as:

$$PLV_{xy} = \left|\left\langle e^{i|\varphi_x(t) - \varphi_y(t)|}\right\rangle\right|$$

where $\varphi_x(t)$ and $\varphi_y(t)$ are the unwrapped phases of the signals *x* and *y* at time *t*. The $\langle . \rangle$ denotes the average over time. The Hilbert transform was used to extract the instantaneous phase of each signal. Unlike MI_{xy}, the h^2 and PLV metrics are normalized so values range from 0 (independent signals) to 1 (fully correlated signals).

Interestingly, loss of consciousness (LOC) (which is a clinical manifestation of epileptic complex seizures) was investigated using non-linear regression analysis applied to intracranial recordings of cortical and subcortical structures in patients with drug-resistant epilepsy, during pre-surgical evaluation (Arthuis et al., 2009; Bonini et al., 2016; Lambert et al., 2012). Results clearly showed a significant increase of long-distance synchronization between recorded brain structures during LOC occurring during seizures and suggest that thalamus and parietal cortices are critical in processing awareness. In addition, the degree of LOC was found to correlate with the amount of hypersynchronization in thalamo-cortical systems, suggesting that excessive synchronization in structures involved in consciousness processing prevents integration of incoming information and contributes to LOC.

Potential relevance for Luminous. *Functional connectivity measures are potentially important in the context of Luminous as they can provide quantified information on brain connectivity, in various conditions either physiological or pathological (sleep, DOCs). Moreover, once a network is defined, its algorithmic complexity or other complex network metrics can be studied, making direct links with information-centric theories of consciousness such as K or IIT. A recurrent question faced by researchers is the choice of the appropriate connectivity measure able to disclose significant differences among different conditions from physiological signals (EEG, MEG or*

BOLD). In previous studies, INSERM partner has already quantified and statistically evaluated the results obtained using different methods on the same simulated data obtained from various models of coupled systems in which a ground-truth on the underlying coupling is available (Wendling et al., 2009). A general conclusion of this comparative study is that none of the tested methods can be considered ideal. On broadband signals, regression methods (linear and nonlinear) exhibited better sensitivity than coherence and phase synchrony methods, what may be interpreted as indicating that interdependence between signals is not entirely determined by a phase relationship. Regarding narrow-band signals sharing a phase relationship, phase synchronization methods performed better, but regression methods yield also correct results, but this was not the case for coherence functions and generalized synchronization methods. In the case of simulated EEG signals with spikes, most methods, except the coherence function and wavelet entropy, could detect the increase of the coupling parameter in the model, but R^2, Hilbert entropy and wavelet mean phase coherence displayed smaller variance of the estimates of the coupling factor. For the Luminous project, we will pay particular attention to the interpretation of results with respect to the applied methods. In addition, we will favour dense-EEG recordings (performed, in particular, in DOCs) and analysis of functional connectivity at the level of cortical sources to avoid drawbacks like volume conductor effects and influence of the choice of the reference (see Appendix B
Description of four classical approaches used to estimate distributed dipole sources).

4.2.1.2 A short overview of methods for assessing effective connectivity

The relationship between two neurophysiological signals goes beyond the notion of *functional connectivity*. Indeed, the question of how the activity of a (neuronal) system may influence that of another system is also essential. To assess this *effective connectivity*, a number of methods based on the concepts of phase and time delay, and on the concept of Granger causality (Granger, 1969) also developed over the two past decades. To estimate the degree of association between two signals and the corresponding time delay in a more general way, the non-linear correlation coefficient h^2 as a function of time shift between two signals *x(t)* and *y(t)* has been introduced by Pijn and colleagues (Pijn et al., 1990). This method has also been shown to give reliable measures for the degree and direction of functional coupling between intracerebral EEG signals both in an animal model of epilepsy (Meeren et al., 2002) and in human epileptic patients (Uva et al., 2005; Wendling et al., 2000). Interestingly, it was also shown that the asymmetry of the measure and the time delay information can be combined in a single quantity named "direction index" which provides reliable information about the information flow (Wendling et al., 2001). Other general methods able to provide information about the direct or indirect influence that a brain neuronal system may exert on another one are, to a large extent, related to Granger causality analysis, that is a signal processing method that provides a quantitative measure of the *causal* association (and by extension of the effective connectivity) between time-series. Besides in econometrics (Granger, 1969), the approach has been applied in various fields including neurophysiology, namely in the analysis of local field potentials (Bernasconi and König, 1999) and it has also inspired applications of related methods in EEG analysis *directed transfer function* (Kamiński and Blinowska, 1991), and *partial directed coherence*, (Baccalá and Sameshima, 2001). Reader may refer to section 4.3.4 for practical details on the computation and usefulness of Granger causality analysis in the context of Luminous.

Potential relevance for Luminous. *Effective connectivity measures are also potentially important in the context of Luminous. Combined with functional measures, they will allow us to characterize connectivity patterns in term of degree and direction, for various states of consciousness.*

4.2.1.3 EEG/MEG source connectivity

Interpretation of connectivity measures from sensor level recordings is not straightforward, as these recordings suffer from a low spatial resolution and are severely corrupted by effects of field spread. To overcome these difficulties, several attempts to apply connectivity methods on the temporal dynamics of brain sources reconstructed from scalp EEG/MEG signals have been reported. Generally speaking, EEG/MEG source connectivity measures involve two main methodological steps: (a) the inverse approach used to estimate the cortical sources and reconstruct their temporal dynamics, (b) the connectivity method used to assess statistically significant functional relationships between the temporal dynamics of sources. These steps are described in *Appendix C*
Steps of cortical sources estimation and methodological issues in EEG source connectivity.

4.2.1.4 EEG/MEG source connectivity and consciousness

Using the keywords “EEG source connectivity” AND “consciousness” in the Pubmed database, we could identify about 10 reports implementing source modeling to investigate brain connectivity patterns in various conditions of consciousness (note that CSD methods are not included here, such as (King et al., 2013a). MEG studies are rarer. Murphy et al. (Murphy et al., 2011) used this technique on high-density electroencephalographic recordings (hd-EEG, 256 electrodes) to investigate the cortical processes underlying propofol anesthesia and compare them to sleep. They found that in loss of consciousness (LOC), EEG slow waves do appear, resembling those of NREM sleep. However, although propofol-induced and natural sleep slow waves share similar cortical origins and propagation (mesial components of the default network), they also display strong differences in the spatial blurring and in the relationship to spindles. Anesthesia-induced LOC was also studied by Boly et al. (Boly et al., 2012) who used source-reconstructed data from frontal and parietal cortices combined with dynamic causal modeling (DCM) to investigate the neural mechanisms mediating spectral changes during normal wakefulness, propofol-induced mild sedation and LOC. Results emphasize the role of corticocortical circuits in the maintenance of consciousness and suggest a direct effect of propofol on cortical dynamics. More recently, the dynamic reorganization of brain functional networks during cognition was analyzed by Bola and Sebel (Bola and Sabel, 2015) using high-density EEG recordings in subjects performing a visual discrimination task. Their findings suggest that dense and clustered connectivity between hub nodes belonging to different modules is the "network fingerprint" of cognition. Authors hypothesize that these reorganization patterns i) facilitate the global integration of information and ii) provide a substrate for a "global workspace" necessary for cognition and consciousness to occur. So far, a large number

of studies mainly focused on phase coupling as a correlate of cortical communication. Interestingly, recent studies suggest that distinct coupling modes may coexist in multi-site communication and participate into cortical functions. This hypothesis was recently studied (Helfrich et al., 2016) based on ambiguous visual stimulation (ambiguous stimuli are commonly used to study the neuronal correlates of consciousness). Source-space connectivity analysis techniques were applied to 128-channel EEGs recorded in participants performing a bistable motion task. The relevance of two different coupling modes (namely phase and envelope coupling) for cortical communication was analyzed. Overall, results suggest that synchronized oscillatory brain activity reflects the processing of sensory input as well as the integrated information across several spatiotemporal scales. They also indicate that the two coupling modes are involved in different cortical computations.

Potential relevance for Luminous. *The above studies confirm that the dynamic interplay of widely distributed, functionally specialized, cortical areas play a central role in consciousness states. With the technical progress achieved in EEG recording systems (in term of number of electrodes and sampling) combined with that gained in signal processing (inverse problem solutions and connectivity measures), it is now possible to estimate the spatio-temporal features of brain networks from non-invasive data with so-called source connectivity approaches. Such approaches present a great potential for the Luminous project as they can provide substantial information about the connectivity of large-scale networks involved in distinct states of consciousness. For corticocortical connectivity, the relevance and applicability of such methods is obvious in sleep studies as well as in DOCs, provided that EEG is recorded on a sufficiently high number of electrodes. For thalamocortical connectivity, the issue is more challenging since subcortical structures in general (and the thalamus in particular) are not represented in the source space. Nevertheless, we cannot rule out the hypothesis that thalamocortical connectivity may induce specific corticocortical dynamic patterns which can be captured by these techniques.*

4.2.2 Network measures based on graph theory

We can also quantify complexity in functional networks derived from, e.g., EEG, MEG or fMRI. Starting from EEG, we can construct, given a few seconds of data, a network (in cortical space) (Ray et al., 2007) from coherence, correlation, mutual information, Granger causality or other related measures (see (Eguíluz et al., 2005) for an example in fMRI of scale-free behavior). This will be typically encoded in the form of an adjacency matrix. Then we can quantify, or try to estimate, the complexity of a graph (e.g., use LZW with the adjacency matrix; Butts, 2001). E.g., a linear approach may include the study of global coherence (the ratio of the largest coherency/cross-spectral matrix eigenvalue to its trace) in different frequencies (Cimenser et al., 2011). From the point of view of K, a simple graph is easy to describe: fully connected or fully disconnected graphs are simple. Sparsely connected graphs in the middle. Those will include small-world or scale- free networks, as it happens.

In more detail, the characteristic path of a network refers to the average shortest path of the network. A small value in the characteristic path results in faster information transfer through the network. Global efficiency estimates how efficiently the vertices

exchange information through the network, concurrently. Local efficiency is defined as the average efficiency of the sub-networks of the neighbors of a node, and can estimate the efficiency of information communication in the network upon the removal of a node. Finally, clustering coefficient estimates the tendency of the nodes in the network to cluster together. Regarding the scale-free network features, Shannon entropy of a network estimates the information presented in the structure of a network, thus, a network with small Shannon entropy is more optimal, and Von Neumann entropy is a measure of regularity in the network.

Potential relevance for Luminous. *Metrics based on graph theory could be used within the Luminous project to explore. Various features can be extracted from small-world or scale-free networks, based on their characteristics. For instance, features such as characteristic path (Watts and Strogatz, 1998), local and global efficiency, and clustering coefficient (Latora and Marchiori, 2001) can be extracted from small-world networks. Regarding scale-free networks, entropy-based features, such as for instance Shannon entropy or von Neumann entropy which emanates from quantum physics and has been recently introduced to network analysis (Passerini and Severini, 2008) can be also extracted.*

4.2.3 Weighted Symbolic Mutual Information (wSMI)

Weighted Symbolic Mutual Information (wSMI) has been proposed in (King et al., 2013b) for characterizing information sharing among brain areas. This measure capitalizes on experimental studies in normal subjects showing that consciously perceived stimuli, relative to subliminal stimuli, lead to a late ignition of frontoparietal networks and to an increased sharing of information in the brain (Dehaene and Changeux, 2011). wSMI aims at quantifying global information sharing through the evaluation of the extent to which two EEG signals present nonrandom joint fluctuations, suggesting that they share information. The measure is based on the definition of symbolic structures in the EEG signals and the characterization of the co-occurrence of the defined symbols in channel pairs for particular time steps. The methodology is parameterized by the symbol length (k) and the time lag τ. wSMI does not simply reflect changes in local entropy or phase-locking and outperforms EEG power spectral analysis in discriminating VS and MCS patients (King et al., 2013b).

Potential relevance for Luminous. *wSMI has been recently included in a comparative analysis of different proposed consciousness metrics within a machine learning approach (Sitt et al., 2014). wSMI has been shown to increase with the consciousness state and to separate vegetative state (VS) from minimally conscious state (MCS)patients (King et al., 2013b). However, these results were obtained at the population level, although detecting consciousness in individual patients should be the actual goal of any consciousness metrics.*

4.2.4 Event Related Potentials (ERPs)

Event-related potentials (ERPs) recorded with EEG as well as Event-related Fields (ERFs) recorded with MEG, are a stereotyped and time-locked electrophysiological response to a specific stimulus (somatosensory, motor or cognitive). Since their discovery in the late 30's (Davis, 1939) they have been used to analyze with high temporal resolution how stimulus information is perceived, processed and integrated. Some late ERP components (see below) have been directly related to the "controlled processing" (Näätänen and Picton, 1986; Schneider and Shiffrin, 1977) of stimuli i.e. its slow, effortful, *conscious* and serial processing as opposed to a fast, efficient, unchecked and parallel processing of the information. Among these components, the P300, CNV and MMN have been more particularly studied in various states of consciousness.

4.2.4.1 P300 component

The P300 component is typically obtained during "odd-ball" paradigms, where the subject is presented with trains of frequent stimuli interrupted rarely and at random times by deviant stimuli. When the subject is asked to detect these deviant stimuli (or target) a large positive wave, the P300, is elicited around the 300ms latency range (Sutton et al., 1965). No P300 can be recorded when the deviant rare stimulus is ignored or undetected (Duncan-Johnson and Donchin, 1977). Originally called P300 this task-relevant component elicited during target stimulus processing has been relabeled P3b after the identification of another wave, called P3a,elicited at earlier latencies (250 ms) when unexpected and highly deviant stimuli are presented even if the subject is not ask to pay attention to it (Squires et al., 1975). This P3a can be considered as part of the so called distracter components that also include other potentials such as the novelty P300 or the no-go potential (Polich, 2007). This P3a has a mostly frontal or fronto-central topographical distribution and is considered to reflect an automatic detection of novelty probably inducing a phenomenon of attention switching. On the contrary, the P3b has a more centro-parietal topography. Its amplitude is proportional to the level of attention of the subject upon presentation of the rare and deviant stimuli. It has been proposed to be related to context updating operations and subsequent memory storage (Polich, 2007).

Historically P3b has been proposed to be in link with the control processing of (rare) stimuli (Posner et al., 1973) and possibly to reflect their conscious perception (Dehaene and Changeux, 2011; Picton, 1992).

Although the P3b component of ERPs is a robust correlate of the subjective report of stimulus detection, as observed in attentional blink paradigms (Sergent et al., 2005) (*Figure 4-5A*), its application as a signature of consciousness has been questioned by several experimental results. Indeed, P3b can be absent despite the engagement of consciousness: if subjects are asked to be aware of both frequent and rare stimuli (pushing a different button for each stimulus type instance) a P3b will be elicited only for rare stimuli and not for the frequent standard ones. Moreover, no P3b (or only a small amplitude one) will be elicited in subjects with absolute pitch when they pay attention to an infrequent sound. (Cote, 2002). Additionally, P3b is reduced or absent when consciously perceived stimuli are not associated with a specific task (task-irrelevant stimuli) (Pitts et al., 2012, 2014). In these two studies however, the proper perception of the task-irrelevant stimuli has been questioned (Rutiku et al., 2015).

Finally the P3b measured during auditory oddball paradigms is often absent in brain-injured conscious patients (Fischer et al., 2010; Höller et al., 2011; Kotchoubey, 2005; Sitt et al., 2014) (*Figure 4-5B*), thus showing an unsatisfactory sensitivity (up to 31%) in identifying minimally conscious patients (King et al., 2013b).

On the other hand P3b can be elicited by stimuli that are not consciously detected, such as in subliminal oddball paradigm (*Figure 4-5C*) (Silverstein et al., 2015). Moreover, a P3b-like potential has been found in 40% of coma patients who are unresponsive, sedated and hypothermic (Tzovara et al., 2015) (*Figure 4-5D*). Therefore, it has been proposed that P3b might rather reflect the consequences of conscious perception (memory update, differences in attention, reporting the content of consciousness) than consciousness itself.

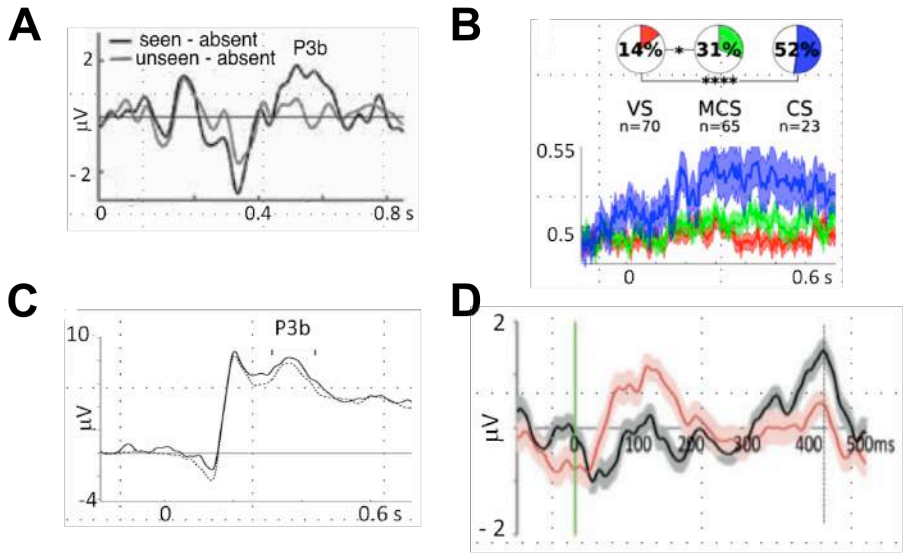

Figure 4-5.Results obtained with P3b in different conditions: attentional blink paradigms (A), auditory oddball paradigm in DOC patients (B), subliminal oddball paradigm in light grey (frequent stimulus) and in dark grey (rare stimulus) (C) and finally in hypothermic coma patients in black (deviant) and in red (standard) (D).

4.2.4.2 Contingent negative variation (CNV)

This component has been first described by Grey Walter in a paradigm where paired stimuli were presented to subject. First stimulus warns the subject and the second stimulus (imperative) necessitates a behavioral response from the subject. Between both stimuli, called the foreperiod (0.5 to 20s), occur (1) brief positive and negative components that can vary with the type of stimuli, and (2) a sustained negative baseline shift called "contingent negative variation" because it varies with the *contingency* between the warning and the imperative stimulus i.e. it depends on the fact that the two stimuli are significantly related from the standpoint of the subject (Walter et al., 1964). During the CNV, the involvement of a wide range of brain areas have been reported (Chennu et al., 2013; Gómez et al., 2001; Nagai et al., 2004) including the frontal, temporal, supplementary motor, cingulate and subcortical areas. The CNV has been directly linked to the stimulus expectation and to the subject's engagement of attention and could mark a conscious state.

4.2.4.3 Mismatch Negativity (MMN)

The MMN is elicited by rare stimuli deviating (in frequency or intensity for example) from the repetitive and frequent standard ones (Näätänen et al., 1978). This component peaks at 100-250s after stimulus onset, has a fronto-central distribution, inverts polarity at the mastoids and is mostly right lateralized. It can be identified by subtracting the response to standard stimuli from that to deviant ones. Because MMN cannot be recorded without the frequent stimuli, it is proposed to reflect a change-detection process based on the memory traces developed by the previous stimulation (Näätänen and Winkler, 1999). This wave is often accompanied by other correlates of attention switching such as the P3a component, the heart-rate deceleration and the skin conductance response (Lyytinen et al., 1992).

MMN is not a necessary condition for a conscious perception and exists when stimuli are not consciously detected. Nevertheless, its amplitude can be modulated by attention (Bekinschtein et al., 2009) and by the conscious state during the wakefulness to sleep transition (Nashida et al., 2000). Moreover its presence in patients with DOC has been shown to be a discriminating predictive factor for favorable outcome even if it failed to taken together, these results globally suggest the limited capacity of MMN to distinguish vegetative from minimally conscious patients (Bekinschtein et al., 2009; Fischer et al., 2004, 2010; Kotchoubey, 2005; Morlet and Fischer, 2014; Naccache et al., 2005; Schnakers et al., 2009; Wijnen et al., 2007).

However, it is quite unknown what is the potential capacity of Mismatch Negativity to detect abstract tones. Therefore, while most of the literature on this ERP state that Mismatch Negativity is a neural correlate of the detection of abstract rules, studies showing this effect have used very simple rule. For example, Saarinen et al. (Saarinen et al., 1992) found that if pairs of two ascending tones were presented as standard tones, MMN appeared when two descending tones were presented. Similar results were found if repeated tones were presented when standard tones were two ascending tones (these results were also replicated using Shepard Tones, (Tervaniemi et al., 1994). Therefore MMN identified the rule “rise”, even when the tones were different. Another study found the appearance of MMN when the rule “the higher the frequency, the louder the intensity” was broken (Paavilainen et al., 2001).

Potential relevance for Luminous. *Mismatch negativity (MMN) and P300b refer to the components of the ERP specifically due to the presence of odd, pattern-breaking stimuli in a sequence . MMN is thought to result from comparison processes between the stimulus and the neural representation of the auditory past maintained in the brain’s sensory memory. It is generated by infrequent simple pattern breaks that do not require attention. Hence it is present even in unconscious states, highlighting the fact that the brain, even in early processing stages, is able to extract patterns. In contrast, the P300b component is thought to reflect a higher-order violation of subjects' expectations of a given rule constructed over a longer time period and has thus been closely linked to conscious access. Recent work suggests that the dissociation between these two neural signatures could discriminate patients in unresponsive wakefulness state from those in conscious or minimally conscious states. Metrics that arise from ERP analysis can be potentially very relevant in the context of Luminous, as they provide an index on the temporal resolution of how information is perceived, processed and integrated. While*

there are no direct link on ERP signatures and the underlying complexity of the neural substrate, ERP metrics are expected to provide important information on the temporal resolution of the brain communication. However, sensitivity of P300b (i.e. the fraction of minimally conscious patients who provide positive test results) is generally low, since it was observed only in 14% (Faugeras et al., 2012) and 31% of MCS patients (King et al., 2013a).

4.2.5 Gamma synchrony

The synchronization of populations of neurons in the gamma range (Crick and Koch, 1990) has been proposed as a correlate of consciousness because this mechanism may account for the binding of multiple stimulus features within a single experience(Singer, 1999). Indeed, two jointly moving visual stimuli produce synchronized neuronal discharges in the low gamma range (30– 70 Hz) in the visual cortex of anaesthetized and awake cats (Gray et al., 1989). Moreover, this synchronization has been shown to be facilitated by attention (Roelfsema et al., 1997) and by stimulation of the reticular formation (Herculano-Houzel et al., 1999; Munk et al., 1996). Also in vivo studies in humans have shown that long-distance gamma synchrony may correlate with visual consciousness (Melloni et al., 2007; Rodriguez et al., 1999). Gamma synchrony in EEG recordings can be measured after 15-Hz high-pass-filtering and time-frequency analysis with the pseudo Wigner-Ville transformation. Then, long-range phase synchrony can be computed for gamma activity as the instantaneous phase value $\varphi_i(f_0, t, k)$, which is a complex number of unit magnitude computed for electrode I, central frequency f_0, time sample t and trial k. Global phase-locking value for an electrode pair i,j computed for all trials k=1,…N can be obtained as:

$$\varphi_{i,j}(f,t) = \frac{\left|\sum_k \varphi_i - \varphi_j\right|}{N}$$

This value is real and ranges between 0 (random phase difference) and 1 (constant phase difference). Normalization for baseline values is crucial for properly compare synchrony values between near and distant electrode pairs.

However, long-range gamma synchronization has been found to correlate with attention, irrespective of actual perception of the stimulus, whereas mid-range gamma synchronization has been found to correlate with stimulus visibility (Wyart and Tallon-Baudry, 2008), working memory retention (Pesaran et al., 2002; Pipa et al., 2009) and temporal expectation (Lima et al., 2011). During early NREM sleep, anaesthesia (Imas et al., 2005; Murphy et al., 2011) or during seizures (Pockett and Holmes, 2009), gamma synchrony can persist or even increase; moreover, it can be present also during exposure to stimuli that induce unconscious emotional responses (Luo et al., 2009). *Figure 4-6A* shows that cross-correlograms computed from neuronal responses recorded 7 millimeters apart in the visual cortex indicate strong gamma band synchronization when a moving bar stimulates the two receptive fields (Gray et al., 1989). Although this and subsequent studies led to the proposal that gamma synchrony may be a correlate of consciousness, it was observed that high-frequency (76-90 Hz) occipital gamma-band activity is modulated by spatial attention both in response to consciously seen and unseen stimuli (Wyart and Tallon-Baudry, 2008) (*Figure 4-6B*). Moreover, stimulus type per se affects gamma activity, as visible grating stimuli elicit robust gamma oscillations (30–80Hz) in visual cortex, while visible noisy patterns do not (Hermes et al., 2015) (*Figure 4-6C*). Finally, when inspecting time series of numbers of channels in synchrony for gamma filtered data (40-46 Hz) (*Figure 4-6D*), gamma synchrony increases as compared to wakefulness (red) during a period of unconsciousness immediately following a generalized tonic–clonic seizure (blue) (Pockett and Holmes, 2009).

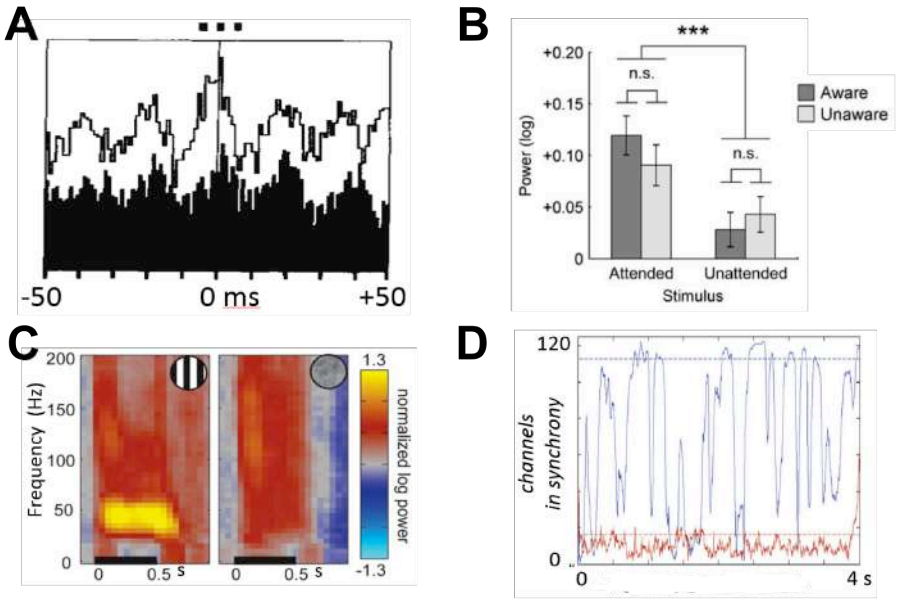

Figure 4-6. Results obtained with gamma synchrony during different conditions: visual stimulation with a moving bar in black (stimulus orientation 180°) and in white (stimulus orientation 0°)(A), spatial attention (gamma power increases with attention irrespectively of awareness) (B), visible grating stimuli and equally visible noisy patterns: only grating stimuli elicit robust gamma oscillations (C), post-critic unconsciousness (in blue) compared to conscious wakefulness (in red) (D).

Potential relevance for Luminous. *In principle, gamma synchrony provides not only a quantitative metric of neural processing, but also mechanistic explanation on how the neural processing is conducted. In this framework, the presence or absence of synchrony is related to the presence of functionally coherent assemblies, a dynamic binding of information that relates to single neuron signaling. However, it has been shown that gamma synchrony can also be observed in the absence of consciousness, turning into a necessary but not sufficient condition for conscious experience.*

4.3 Information- integration-based metrics

4.3.2 Neural Complexity (C_N)

Neural complexity (C_N) expresses the extent to which a system S is both dynamically segregated (i.e. small subsets of the system tend to behave independently) and dynamically integrated (i.e. large subsets of the system tend to behave coherently) (Tononi and Edelman, 1998; Tononi et al., 1994). C_N is high if each subset can take on many different states and if these states make a difference to the rest of the system. From a mathematical perspective, C_N of a system S is equal to the sum of the ensemble average mutual information (MI) across all bipartitions of the system (Tononi et al., 1994). MI between two subsets A and B is computed as:

$$MI(A;B) = H(A) + H(B) - H(AB)$$

where H(●) is the informational entropy, i.e., the overall degree of statistical independence. Under Gaussian assumptions, the informational entropy of a system S can be calculated analytically from the covariance matrix (COV(●)) relating the responses of the elements of the system as follows:

$$H(S) = 0.5ln((2\pi e)^n|COV(S)|)$$

where n is the number of elements of the system and $|\cdot|$ is the determinant of the covariance matrix. Given the above, C_N of a system S is computed using the following formula:

$$C_N(S) = \sum_{k=1}^{n_t/2} \langle MI(S_j^k; S - S_j^k) \rangle$$

where k is the subset size, n_t is the total number of subsets of size k, S^k_j is the j-th bipartition of size k, $\langle \cdot \rangle$ indicates the ensemble average across index j.

C_N reflects the explicit exchange of signals that takes place either within the isolated system or in a behaving system during interaction with an external environment as an embedded and embodied neural network (Seth and Edelman, 2004). Initially, C_N was used to characterize a neural system isolated from the environment; then, it has been extended to characterize the change in C_N that occurs after a neural system receives signals from the environment (matching complexity, C_M) (Tononi et al., 1996). C_M is low when the intrinsic connectivity of a system is randomly organized, while is high when the intrinsic correlations that are enhanced by sensory input are differentially amplified.

Precise calculation of C_N requires the evaluation of mutual information across all possible bipartitions, which can become computationally prohibitive for large systems. However, C_N can be approximated by considering only those bipartitions that consist in one single element against all of the remaining elements (Tononi et al., 1998c). A disadvantage of C_N and its approximation is that they do not reflect causal interactions because C_N is based on mutual information, which is a symmetric quantity.

Potential relevance for Luminous. *Neural Complexity is practically impossible to compute, when simplified version where tested they yielded inconclusive results. Thus, the current formulation of* C_N *may not suite the scope of the project.*

4.3.3 Integrated Information (Φ)

The measure Φ represents an evolution and a generalization of C_N and aims at assessing integrated information by a causal and intrinsic perspective (Tononi, 2004). In particular, the symbol Φ indicates that the information "I" is integrated within a single entity "O".

Φ can be theoretically computed (i) by identifying all possible bipartitions of a system (any A|B) and, for each of them, (ii) by measuring how differentiated are the responses of B when A is perturbed by replacing its output by uncorrelated noise (i.e., maximal entropy Hmax). This is the effective information (EI) between A and B, i.e. the causal influence of A on B, and is expressed by the following formula:

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

where MI(●) is the mutual information [2].

Given that EI(A→B) and EI(B→A) are not necessarily equal, one can define:

$EI(A \leftrightarrow B) = EI(A \rightarrow B) + EI(B \rightarrow A)$

that represents a measure of the repertoire of all possible causal effects of A on B and of B on A.

In neural terms, we evaluate how differentiated is the repertoire of firing patterns produced in B by all possible combinations of firing patterns output from A. EI(A→B) will be high if the connections between A and B are strong and specialized (i.e. different firing patterns from A will induce different effects in B), while will be low or zero if A and B are connected such that different outputs from A produce scarce or stereotypical effects on B.

A subset of a system cannot integrate any information *per se* if there is a partition A|B such that EI(A↔B) = 0: in this case, A and B are causally independent. Similarly, a subset can integrate little information if there is a partition such that EI(A↔B) is low: this EI is the limiting factor on the subset's integrated information capacity. Therefore the integrated information capacity of a subset is bounded by the bipartition(s) of S for which EI(A↔B) reaches a minimum (the informational "weakest link"). EI(A↔B) is by definition bounded by the maximum entropy available to A or B. In order to find the minimum EI(A↔B) across bipartitions, it needs to be normalized by min{Hmax(A); Hmax(B)}, i.e. the maximum information capacity for each bipartition. Therefore, Φ(S) is the non-normalized value of EI(A↔B) for the minimum information bipartition (MIB).

Within a system it is possible to identify complexes, i.e. subsets that have Φ>0 and are not included within larger subsets with higher Φ. Among the complexes that are identified within a given system, the one with maximum Φ is called the main complex. The spatial and temporal scales used for evaluating a system influence the definition of its state: therefore, Φ will be maximum at a specific spatiotemporal grain size.

---

[2] An extension of this can be considered by using mutual algorithmic information rather than Shannon's MI.

As compared to C_N, Φ reflects causal interactions because is computed from a directional version of mutual information, i.e. effective information. However, Φ cannot be measured for any nontrivial real-world system because (i) it is impracticable to replace the outputs of arbitrary subsets of complex real neural systems with uncorrelated noise and (however, here we note that computational models of the brain can be probed using such perturbative techniques) (ii) the number of partitions to be evaluated grows with a factorial law as the size of the network increases (same for C_N computation). Moreover, a reliable definition of an approximated measure has not been developed so far (Tononi, 2004). However, for simple systems with a limited number of elements, there is a Python-based library (PyPhi) to compute integrated information and the associated quantities and objects (https://pypi.python.org/pypi/pyphi).

Potential relevance for Luminous. *Integrated information can be explored in a computational model of the brain. Empirically, its full implementation is currently prohibitive, although simplified measurements inspired by Φ can be developed.*

4.3.4 Causal Density (C_d)

Causal density (C_d) is a measure of causal interactivity that captures dynamical heterogeneity among network elements (differentiation) as well as their global dynamical integration (Seth, 2005). Specifically, C_d is a measure of the fraction of interactions among neuronal elements that are causally significant and ranges between [0,1]. This measure is based on Granger causality (Granger, 1969), that can be simplified as follows: if one signal causes another signal, then past values of the first signal should be able to predict the second signal without any knowledge of past values of the second signal itself. Granger causality can be measured by applying multivariate linear regression models (Hamilton, 1994) as follows:

$$\begin{cases} x_1(t) = \sum_{j=1}^{p} A_{11,j} x_1(t-j) + \sum_{j=1}^{p} A_{12,j} x_2(t-j) + E_1(t) \\ x_2(t) = \sum_{j=1}^{p} A_{21,j} x_1(t-j) + \sum_{j=1}^{p} A_{22,j} x_2(t-j) + E_2(t) \end{cases}$$

where x_1 and x_2 are two time series of length T, p is the maximum number of lagged observations included in the model (the model order, $p < T$), the matrix A contains the model coefficients and E_1 and E_2 are the prediction errors for each time series. If the variance of E_1 (or E_2) is reduced by including x_2 (or x_1) terms in the first (or second) equation, then x_2 (or x_1) ‘‘Granger-causes’’ x_1 (or x_2). This relationship can be tested with an F test of the null hypothesis that $A_{12} = 0$, assuming that the covariance of x_1 and x_2 is stationary. The logarithm of F statistic represents the magnitude of the Granger causality interaction (Geweke, 1982). This approach can be readily extended to n variables. The C_d of a system is computed as:

$$C_d = \alpha/(n(n-1))$$

where α is the total number of significant causal interactions and n(n-1) is the total number of directed edges in a fully connected network with n nodes, excluding self-connections.

C_d is high when the elements of a system are both globally coordinated in their activity (can be used for predicting each other's activity) and at the same time dynamically distinct (different elements contribute in different ways to these predictions). Causal density cannot be obtained from network anatomy alone but requires also the analysis of time series representing the dynamic activity of neurons. C_d takes into account all causal interactions within a system, while φ depends on the interactions across a specific single bipartition.

Potential relevance for Luminous. *Although the computational burden for C_N and φ is higher, also in case of C_d computation, the estimation of multivariate regression models becomes difficult with an increasing number of network elements. For instance, the total number of parameters for a network of n elements grows as pn^2 and the number of parameters to be estimated for any single time series grows linearly as pn. In order to overcome this issue, Bayesian methods have been proposed to reduce the number of model parameters by introducing prior constraints on significant interactions (Zellner, 1996). Concerning neural systems, it is possible to resort to known neuroanatomical constraints.*

4.3.5 Coalition entropy measures (ACE and SCE)

Recently, Schartner and colleagues (Schartner et al., 2015) proposed two novel candidate measures for consciousness, the amplitude-coalition entropy (ACE) and the synchrony-coalition entropy (SCE). These authors shown that these two novel metrics robustly distinguished loss of consciousness due to propofol anaesthesia from wakefulness on the broadband signal, resulting in higher values for wakefulness as compared to LOC across subjects, a range of segment lengths, and number and location of electrodes. They also demonstrated a correlation between level of sedation and both complexity measures (i.e. ACE and SCE but they also employed LZW) during propofol-induced general anaesthesia.

These two complexity metrics were variants of the coalition entropy previously introduced by Shanahan et al (Shanahan, 2010).

In particular, ACE reflects the entropy over time of the constitution of the set (i.e. coalition) of the most active channels. From a computation point of view, ACE is similar to LZW for the binarization scheme to classify active channels and for the normalization procedure. From a more conceptual point of view, ACE is similar to LZW since this metric quantifies the variability of the EEG activity in space and time.

On the other hand, the novel SCE reflects the entropy over time of the constitution of the set of synchronous channels. SCE is conceptually different from ACE because it quantifies variability in the relationships between pairs of channels and measures the diversity, over time, of the coalition of channels that are in synchrony rather than active.

Channels were considered in synchrony at time t if the absolute value of the difference between their instantaneous Hilbert phases is less than 0.8 radians (approximately 45 degrees).

The coalition time-series $\boldsymbol{\Psi}_t^{(i)}$ by $\Psi_t^{(i)}$ was defined considering the value 1 if channels (i.e. *i* and *j*) are synchronised at time *t* and taking the value 0 otherwise.
The coalition entropy $SCE^{(i)}$ with respect to channel *i* is the entropy over columns of the matrix $\boldsymbol{\Psi}_t^{(i)}$ (over time) containing all synchrony time series for channel *i*, normalized as a proportion of its maximum possible value N:

$$SCE^{(i)} = -\frac{1}{N}\sum_{\psi} p(\boldsymbol{\Psi}_t^{(i)} = \psi)\log p\ (\boldsymbol{\Psi}_t^{(i)} = \psi)$$

The overall coalition entropy **SCE** is the mean value of $SCE^{(i)}$ across channels.
Potential relevance for Luminous. *These theory-driven metrics reflect complexity, interpreted as co-existing differentiation and integration. In this sense, ACE and SCE captures differentiation in the sense of signal diversity (variation of amplitude and synchrony coalitions) over time; while integration was indexed by the variations over time to the set of active or synchronous channels*
ACE and SCE are easy to compute and do not need a direct stimulation of the brain (e.g. perturbing as TMS) since they are based on the spontaneous steady-state EEG data. Results obtained on propofol-induced anesthesia are promising and definitely prompt a further investigation on different unconsciousness states such as the NREM sleep and the vegetative state.

4.3.6 Perturbational Complexity Index (PCI)

The Perturbational Complexity Index (PCI) has been defined as the algorithmic complexity of a system's deterministic response to a perturbation (Casali et al., 2013). As such, PCI captures the complexity of the neural activity patterns that are caused by direct perturbation, reproducible and virtually insensitive to random processes. It has been proposed that consciousness theoretically depends on the ability of distributed regions of the brain to interact through divergent cortico-cortical and cortico-thalamo-cortical connections (Tononi, 2004). Inspired by this theoretical framework, PCI can be computed on brain responses to direct cortical perturbation. Experimentally, the combination of high-density electroencephalography and transcranial magnetic stimulation (TMS/EEG) allows to (i) directly and non-invasively activate a bundle of neuronal axons that originate from the cortical surface and (ii) to record with optimal spatiotemporal resolution the dynamic interactions among neuronal elements that are elicited by this perturbation (Hallett, 2000; Virtanen et al., 1999).Once the cortico-cortical TMS-evoked potentials are recorded from the scalp and the corresponding cortical current density distribution in the brain has been obtained by applying source modelling (Berg and Scherg, 1994; Friston et al., 2006; Mattout et al., 2006; Phillips et al., 2005; Zhang, 1995), PCI computation involves pre-processing steps, which are (i) extracting the deterministic patterns of cortical activation by statistical analysis and (ii) estimating their information content by applying measures of algorithmic complexity. Non-parametric bootstrap-based statistical procedure can be effectively applied to estimate the statistically significant ($\alpha = 0.01$) deterministic responses of the brain to TMS (Lv et al., 2007; Pantazis et al., 2005). In this way, the spatiotemporal distribution of significant sources can be represented by a binary matrix SS(●), whose element (x,t) is 1 for significant sources (x) and time samples (t) and 0 otherwise. In order to compute PCI, it is necessary to approximate algorithmic complexity with the Lempel-Ziv complexity (c_L) method (Lempel and Ziv, 1976). By applying a modified version of (Kaspar and Schuster, 1987), the Lempel-Ziv complexity of the binary matrix SS(x,t) can be computed by running the algorithm through the several columns of the matrix

while keeping track of patterns progressively encountered. The asymptotic behavior of c_L for random strings of length L is:

$$LH(L) / \log_2 L$$

where H(L) is the source entropy

$$H(L)=-p_1\log_2 p_1-(1-p_1)\log_2(1-p_1)$$

and p_1 is the fraction of '1' contained in the binary string. Since maximum complexity is extremely sensitive to source entropy, PCI is computing by normalizing the Lempel-Ziv complexity of matrix SS(x,t) as follows:

$$PCI = \overline{c_L} = c_L log_2 L / \big(LH(L)\big)$$

where L is the total number of elements in matrix SS(x,t). PCI ranges between [0,1] and, asymptotically in L, $\overline{c_L} = 1$ for strictly random sequences. This normalization yields a complexity measure that is minimally dependent on the total amount of significant activity and maximally dependent on the formation of patterns in the data. Furthermore, PCI increases with the number of different spatial patterns of binary significant cortical activity occurring in a given time sample that do not occur in previous samples. It is possible to obtain the time course of PCI by iteratively estimating the algorithm complexity of the first column, of the first two columns, of the first three columns … of matrix SS(x,t). In principle, the algorithmic complexity depends on the ordination of SS rows (sources) and therefore the actual complexity of SS(x,t) should be the minimal one across all permutations of sources. Without searching for all possible permutations, it is more convenient to approximate the optimal spatial ordination with lower complexity by sorting sources by their number of significant samples. In order to avoid numerical instabilities related to normalization by an entropy value close to zero, PCI was calculated only if p_1 was greater than the rate of false positives of the statistics (1%, α=0.01), corresponding to an entropy H > 0.08.

Potential relevance for Luminous. *In the context of Luminous, PCI has demonstrated its clinical relevance to evaluate accurately the degree of consciousness* (Casarotto et al., 2016). *Given its reliability, we propose that novel promising metrics of consciousness developed in the context of the Luminous project should be tested against PCI to compare their performances.*

4.4 Machine Learning

Machine learning algorithms such as Support Vector Machines (SVM) which has been already commonly used in EEG signals, e.g. (Güler and Ubeyli, 2007), Echo State Networks (Jaeger, 2001), or Random Forests (Breiman, 2001) will be also used for classification purposes, in order to automatically distinguish across various states of consciousness. Although these classifiers have been extensively used in other fields, it is still a challenging issue to automatically discriminate across disorders of consciousness (Noirhomme et al., 2015). Moreover, feature and/or decision fusion of the most relevant features for consciousness estimation will be carried out, in order to combine the various features that we will develop throughout the project and to automatically distinguish across various states of consciousness in an optimized way.

For instance, Sitt et. al. (Sitt et al., 2014) used a linear SVM classifier to discriminate between MCS and UWS, and revealed that low-frequency EEG power, EEG complexity, and information exchange when combined a low an automatic classification of a patient's state of consciousness with an area under curve (AUC) of 78%. Also, in (Höller et al., 2014) the authors applied SVM classification between MCS and UWS patients, and healthy controls, and revealed that features such as partial coherence, directed transfer function, and generalized partial directed coherence yielded accuracies significantly higher than chance. Simpler classifiers have been also used for automatic recognition of diseases of consciousness. For instance a linear discriminant analysis (LDA) classifier was used in a nested block-wise cross-validation scheme, to discriminate across various diseases of consciousness through complex mental imagery and passive feet movements tasks (Horki et al., 2014). Although various classification approaches have been already used in consciousness research, advanced, more recent classification approaches, such as deep neural networks using autoencoders, echo state networks or random forests still lack attention. Due to the complicated nature of consciousness and to the many different features and their properties related to it, we believe that advanced machine learning approaches that can learn the structure of complex data can reveal additional information about consciousness.

Potential relevance for Luminous. *Machine learning can be potentially very relevant in the context of Luminous as a means of compression of the original data space that reveals the most informative representation of the data, which can be further compressed using methods previously described, such as LZW. Moreover, machine learning can be also used in our context as a means of classification across various states of consciousness.*

4.5 Circularity problem

Several neurophysiological measures have been proposed to measure consciousness; however, at the individual level, these measures must be interpreted with caution because their accuracy cannot be reliably estimated. The performances of a measure are generally tested on a population of individuals who are known either to have or have not a specific characteristic of interest. In this case, sensitivity and specificity of the measure can be derived from the number of individuals whose categorization corresponds with the true state-of-affairs. A measure is considered a gold standard when it reliably identifies the highest number of true positives and true negatives results in a population.

The assessment of consciousness in patients is affected by a general problem of logic circularity, due to the lack of a ground truth, i.e. the actual presence or absence of consciousness (Harrison and Connolly, 2013). Currently, the best available diagnostic tool is based on a standardized behavioral evaluation (i.e. Coma Recovery Scale Revised, CRS-R (Giacino et al., 2004), which is typically repeated over time in order to detect minimal signs of consciousness even when behavioral responsiveness fluctuates. However, the lack of responsiveness does not always correspond to a lack of consciousness because, by definition, subjects may be unable to respond to stimuli while still having conscious experiences (Fernández-Espejo and Owen, 2013; Laureys and Schiff, 2012). This may happen because of motor or executive functions impairments (Fernández-Espejo et al., 2015; Schiff, 2010) and/or because of sensory disconnection from the environment (Sanders et al., 2012). For example, a patient can be covertly conscious but at the same time behaviorally unresponsive: in this case, even if the neurophysiological measure correctly provides a positive result, it would be paradoxically considered a false positive, thus underestimating the performance of the measure to detect the actual presence of consciousness. Despite its intrinsic weaknesses, behavioral clinical assessment is often referred to as an absolute reference to validate more objective measures of consciousness. As such, it is not possible to reliably compute the actual accuracy of any measure of consciousness because of the lack of a veridical benchmark of the true state-of-affairs (Peterson et al., 2015).

In a recent work (Casarotto et al., 2016), this issue has been tackled by first applying a candidate measure of consciousness (PCI) in a large benchmark population of 150 subjects who could confirm the presence or absence of conscious experience through immediate or delayed reports. This benchmark population included: (1) healthy awake subjects of different age (range 18-80 years); (2) conscious brain-injured patients who were awake and able to communicate; (3) unresponsive subjects who reported no conscious experience upon awakening from NREM (non rapid-eye-movement) sleep; (4) unresponsive subjects who reported no conscious experience upon awakening from midazolam, xenon and propofol anesthesia; (5) subjects who were disconnected and unresponsive during rapid-eye-movement (REM) sleep and ketamine anesthesia but retrospectively reported having had vivid conscious experiences upon awakening. PCI has been validated in this benchmark population by considering subjects' reports as the provisional gold standard for assessing consciousness (no report = unconscious condition; immediate or delayed report = conscious condition). Receiver operating characteristic (ROC) curve analysis allowed deriving an optimal operational cutoff able

to distinguish between conscious and unconscious conditions with 100% sensitivity and specificity. Then, this independently validated cutoff could be applied to a population of DOC patients (38 MCS and 43 VS) in order to objectively evaluate the capability of PCI in detecting the potentiality for consciousness in severely brain-injured patients with disorders of consciousness. Results showed that PCI (i) had a sensitivity of 94.7% in detecting MCS patients and (ii) provided a stratification of VS patients into three subgroups of different complexity that might have significant therapeutic implications.

5 Conclusions

5.1 Overview of the documents and guiding principles for novel metrics

At the outset of this document, we have reviewed four theoretical models of consciousness that have implications for the development of empirical metrics. This overview revealed common features as well as specificities. By a practical standpoint e found that one conceptually useful way to relate these models is to consider whether a given model explicitly put more emphasis on integration processes or on the information content of neuronal activity. While we found relative differences to this regard, it is clear that a common denominator should include both integration and differentiation processes.

This analysis was somehow paralleled by an overview of the currently available metrics of consciousness. Also in this case, we found a common thread related to the concepts of neuronal integration and differentiation. Each metric appeared to put a more or less explicit emphasis on the concept of integration, on differentiation or on both. In search for taxonomy, we have chosen to group existing empirical metrics accordingly, although we are aware that this approach is necessarily very schematic. In fact, as it is the case for theoretical models, each measure could be roughly laid on a continuous plane between the two axes of integration and differentiation.

Reviewing the empirical literature (part of this review carried out during the Luminous project was published in (Koch et al., 2016) suggests that the EEG remains a fundamental clinical tool for discriminating between conscious and unconscious individuals (Forgacs et al., 2014). Such an evaluation typically includes qualitative features of the spontaneous EEG (such as activation) and how the EEG responds to perturbations (such as eyes opening) that capture the degree of spatio-temporal differentiation of electrical activity in the brain (Synek, 1988). This classic notion of the "activated EEG", often overlooked in the search for NCC, suggests that a large repertoire of neural activity patterns (differentiation) is important for consciousness (Tononi, 2012). Indeed, recent studies support this view. For example, in rats, the number of unique fMRI BOLD patterns increases upon recovery of consciousness from desflurane anesthesia (Hudetz et al., 2015). The repertoire of fMRI functional connectivity configurations is greater during wakefulness than during propofol anesthesia in monkeys (Barttfeld et al., 2015). Likewise, electrocortical dynamics become more stable upon loss of consciousness, regardless of anesthetic-specific effects on activity (Solovey et al., 2015). In humans, measuring the entropy of brain activity is used to assess the depth of anesthesia (Sigl and Chamoun, 1994) and provides a useful prognostic index of recovery of consciousness in vegetative patients (Gosseries et al., 2011; Sarà et al., 2011).

It is generally recognized that consciousness also requires an integrated neural substrate.[1] This notion is supported by empirical studies. Thus, fMRI studies of whole-

brain functional connectivity show that integration decreases and modularity increases when consciousness is lost during sleep (Boly et al., 2012; Tagliazucchi and Laufs, 2014), anesthesia (Achard et al., 2012) and coma (Monti et al., 2013).On a finer time-scale, similar conclusions have been obtained by measuring functional connectivity using EEG, especially in the alpha and the theta range (Chennu et al., 2014; King et al., 2013b; Marinazzo et al., 2014).

Reappraising the classic EEG, its relationships with the theoretical notions of integration and differentiation, as well as its performance – which is comparable to the one of recently quantitative metrics – is a fundamental step in the context of Luminous.

On the other hand, a critical analysis of the literature, suggests that electrophysiological measures that tend to capture neural integration alone or neural differentiation alone may lack accuracy. For example, indices of cortical integration (such as EEG coherence and Granger causality) can increase in conditions in which consciousness is lost, such as during propofol anesthesia (Supp et al., 2011) or generalized seizures (Arthuis et al., 2009), while EEG measures of differentiation (such as the bispectral index and spectral entropy) are only useful at the group level owing to wide variations across subjects and to the contamination of the EEG signal by noisy (high-entropy) sources, such as muscle activity and environmental noise (Kaskinoro et al., 2011; Sitt et al., 2014).

In order to overcome these pitfalls, one may devise practical ways to explicitly assess the joint presence of integration and differentiation based on the electrical activity of corticothalamic networks. However, this task is practically challenging; as illustrated in *Figure 5-1*, when typical measures of integration (such as synchrony, phase-locking or coherence) are high, typical differentiation indices (such as entropy and algorithmic complexity) tend to be low, and vice versa. A notable exception is represented by a recently introduced set of spontaneous EEG-based measures which were recently tested under propofol anesthesia (Schartner et al., 2015 already described in section 4.3.4).

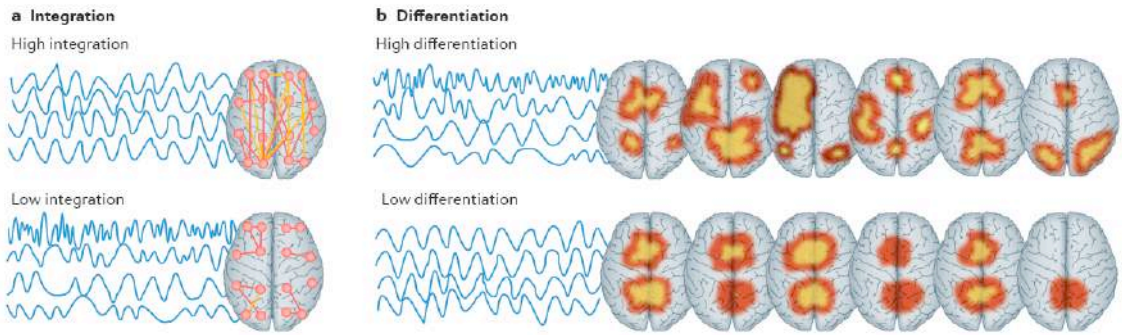

Figure 5-1 Assessing neural differentiation and integration. a) Schematic diagram showing two idealized time series of spontaneous brain activity (electroencephalography (EEG)) and the corresponding maps of cortical connectivity. Integration, as measured by indices of functional connectivity, tends to be high when time series are highly correlated (top panel) and low when time series are not highly correlated (bottom panel). b) The same two time series are shown together with the corresponding maps of spatiotemporal variability in brain activity. Differentiation, here reflected in the difference between subsequent maps of cortical activity, tends to be high (top panel) when the time series are not highly correlated (maximum for random time series) and low when they are highly correlated (bottom panel).

One established approach to simultaneously quantifying integration and differentiation in the human brain is represented by the perturbational complexity index (PCI). As

already described above, calculating the PCI involves perturbing the brain with transcranial magnetic stimulation (TMS), recording the results of EEG to detect the pattern of causal cortical interactions engaged by the TMS perturbation (integration) and compressing this pattern to calculate its spatiotemporal variability with algorithmic complexity measures (differentiation). Responses that are both integrated and differentiated are less compressible, resulting in high PCI values. By contrast, local (low integration) or stereotypical (low differentiation) EEG responses to TMS can be effectively compressed, yielding low PCI values. Unlike measures of differentiation of spontaneous activity, PCI evaluates the deterministic responses of cortex to perturbations and is therefore largely insensitive to random processes or to locally generated patterns that are not genuinely integrated. Furthermore, unlike measures of integration that rely on widespread neural synchronization, the PCI is low when neural activations are spatially extended but undifferentiated, as is often the case during anesthesia and generalized seizures. Finally, because it bypasses sensory pathways, TMS can be used to assess consciousness also when subjects are disconnected from sensory inputs and motor outputs, such as during REM sleep and ketamine anesthesia (Sarasso et al., 2015).

A recent work (Casarotto et al., 2016), which was finalized during the Luminous project, demonstrated maximum accuracy (100%) of this index in discriminating between consciousness and unconsciousness in a large benchmark population (N=150) and an unprecedented sensitivity (n=94.7%) in detecting minimally conscious patients. This result suggests that gauging simultaneously integration and differentiation provides an edge in the practical assessment of consciousness.

However, the dependency of PCI on a relatively complex experimental apparatus, i.e. combining TMS and HD-EEG, may limit its general clinical applicability. More specifically, it is worth noting that the PCI approach, as currently formulated, may be unpractical (or unfeasible) in some of the conditions considered in the Luminous project. Therefore it is important to seek for alternatives offering a good characterization with a more general application scope. In addition, since one of the aims of Luminous is to use metrics of consciousness to guide neuromodulation in a closed-loop fashion, their computation must be relatively fast.

Essentially, as suggested by the overview of models and metrics reported in the present document, this search should be guided by few basic principles. Thus, the novel metric should:

- tend to gauging the integration and the differentiation of neural activity at once;
- be generally applicable, fast and easy to compute;
- be validated in a benchmark cohort of subjects who are able to report about their state of consciousness.

5.2 Towards novel metrics of consciousness

In the final section of this document, we propose a list of possible approaches that may satisfy the basic requirements outlined above. We will first outline a set of possible approaches requiring perturbations and a more complicated set-up and then the ones that can be simply based on the analysis of the spontaneous EEG.

5.2.1 Perturbation-based approaches

5.2.1.1 A simpler and faster computation of PCI

The computation of PCI must be performed off-line because it requires computationally demanding procedures (i) to estimate cortical current density from scalp potentials by solving the inverse problem and (ii) to detect the spatiotemporal distribution of significant current sources by means of data-driven statistics. However, in order to use PCI on-line as a marker of consciousness to inform closed-loop neuromodulation approaches it is crucial to simplify the computation of this index possibly at the scalp level and using a limited number of trials. Thus, we will explore the minimal requirements in terms of number of trials and computational load that is needed to achieve reliable results. Specifically, we will explore the possibility of estimating complexity at the scalp level and we will search for the minimum number of trials that allow a simplified statistical analysis, without losing precision as compared to the standard offline computation. In parallel, we will test the possibility of computing indices of complexity starting from peripheral stimulations collected during longitudinal measurements as well as during sleep in control subjects.

5.2.1.2 A quantification of cortical bistability

Bistability is a neurophysiological mechanism that occurs in various physiological, pharmacological or pathological conditions, such as an increase of K+ conductances, an alteration of the excitation/inhibition balance in favor of inhibition, and cortical deafferentation. Bistability involves a generalized loss of both selectivity and effectiveness, that result in a breakdown of differentiation and integration. Therefore, bistability might be crucial also when thalamocortical integrated information and consciousness are impaired despite preserved neuronal activity, such as epilepsy, anesthesia and the vegetative state/unresponsive wakefulness syndrome. An experimental evidence consistent with this account has been recently reported using intracranial stimulation and recordings in patients with epilepsy (Pigorini et al., 2015). During wakefulness, electrical stimulation triggers a long-lasting chain of deterministic effects (i.e. sustained phase-locked activations in distant cortical targets). On the contrary, during slow wave sleep, the same input induces a slow wave associated with a cortical down-state (suppression of power >20 Hz, reflecting neuronal silence), that is followed by a resumption of cortical activity to wakefulness-like levels without any phase-locking to the stimulus. Thus, cortical circuits, upon receiving an input, tend to respond briefly, then hush and forget, indicating a break in the cause-effect chain. This

study suggests that bistability may be one of the mechanisms that reduce the brain's capacity to integrate information during slow wave sleep. In order to evaluate bistability, it is necessary to quantify (i) the amplitude of low frequency oscillations - SWa (absolute value of the average of 4Hz low-pass filtered trials and bootstrap-based statistical analysis to detect significant time points with respect to baseline), (ii) the suppression of high frequency power > 20 Hz - PWR (Wavelet-based time-frequency decomposition, normalized absolute spectra and bootstrap-based statistical analysis to detect significant time-frequency points with respect to baseline) and (iii) the PLF (Palva et al., 2005) (band-pass filtering between 8-100 Hz of single trials, absolute value of the average of the Hilbert Transform of all single trials and statistical analysis based on Rayleigh distribution to detect significant PLF values with respect to baseline at $\alpha<0.05$) (*Figure 5-2*).

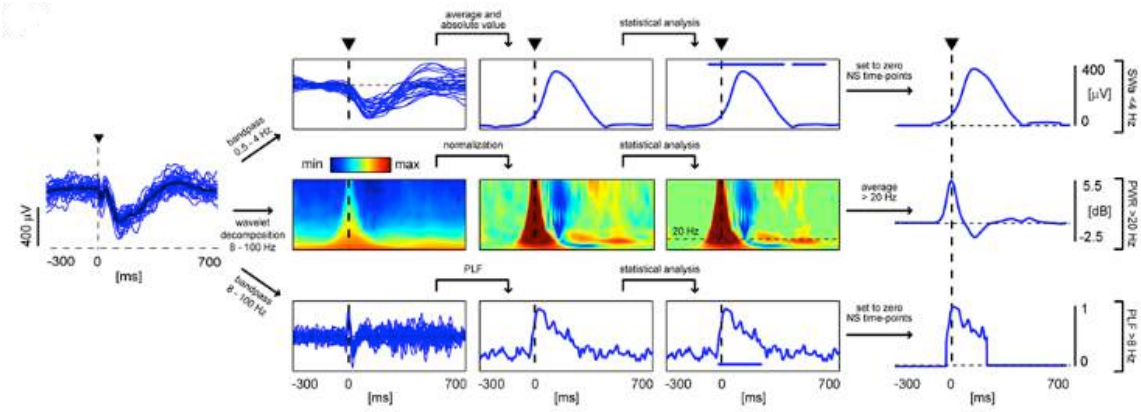

Figure 5-2. Methodological rationale and procedure for evaluating bistability. Single electrical responses to stimulation are analyzed to compute (1) the amplitude of the low frequency components (4Hz low-pass filtering, averaging and rectification, zero-ing of non-significant time points according to bootstrap statistics at $p < 0.01$*); (2) the suppression of high frequency power (Wavelet decomposition, normalization of absolute spectra for the baseline, zero-ing of non-significant time points according to bootstrap statistics at* $p < 0.05$*); (3) the phase-locking factor (PLF) (8-100 Hz band pass filtering, averaging and rectification of the Hilbert Transform, zero-ing of non-significant time points according to statistical analysis at* $p < 0.05$*). [modified from (Pigorini et al., 2015)].*

While this quantification of bistability has been successfully performed based on intracranial stimulation and recording data, its applicability to non-invasive measurements (such as TMS and EEG) has not been systematically assessed so far. Thus, in the course of the project we will test the feasibility and the accuracy of bistability measures on TMS/EEG data acquired in both healthy controls and DOC patients.

5.2.1.3 Assessing integration and differentiation with tACS+EEG,

Another approach to jointly assess differentiation and integration may be represented by the combination of transcranial alternating current stimulation (tACS) and EEG. We will test this novel idea during the project. We will build up on a recent work, which pioneered a novel approach for simultaneous tACS-EEG recordings and successfully separate stimulation artifacts from ongoing and event-related cortical activity (Helfrich et al., 2014). The results of this work showed that the externally applied electric field does directly influence cortical oscillators in a frequency-specific manner. Specifically,

this work demonstrated that (1) 10 Hz tACS increased oscillatory power in the alpha band; (2) the alpha power increase was based on synchronization to the external driving force as assessed by phase-locking analysis (Thut et al., 2011). (3) 10 Hz tACS transiently shifts the individual alpha peak toward 10 Hz.

Based on these premises, it is possible to envisage a rather simple procedure by which integration and differentiation are jointly assessed by means of tACS and EEG. The protocol would involve applying tACS at different frequencies according to a principle similar to the on employed when eliciting sensory-evoked steady state potentials (Vialatte et al., 2010). Based on this data, integration can be then assessed as the spread over the cortex of the synchronization enforced by tACS at each frequency, whereas differentiation can be simultaneously assessed by measuring the difference between the cortical responses elicited by tACS at different frequencies. Responses to tACS that are specific for the stimulation frequency but local will indicate that differentiation is high whereas integration is low. Responses to tACS that spread globally over the cortex but they do so in an unspecific manner (different driving tACS frequencies resulting in cortical responses with similar frequency) will indicate that integration is high, whereas differentiation is low. Instead, responses to tACS that are both widespread and highly specific for the stimulation frequency will reflect an optimal balance between integration and differentiation. A similar approach may be also applied in a simpler protocol where tACS is injected in bursts, and the EEG recorded in the inter-stimulus periods. If this approach is successful it may represent a simpler way to assess integration and differentiation by means of perturbation (tACS may be more practical than TMS at the bedside) and a straightforward way to implement closed-loop tACS protocols aimed at restoring consciousness.

5.2.2 Spontaneous EEG–based approaches

5.2.2.1 Modularity of networks identified from EEG source connectivity methods

Algorithms based on graph theory can be applied to characterize and quantify the topological properties of brain networks identified from neuroimaging data. This network-based analysis has been shown to reveal various properties of studied networks, such as small-worldness (Bassett and Bullmore, 2006), hubs (Sporns, 2011), rich-club (van den Heuvel and Sporns, 2011) and modularity (Sporns and Betzel, 2016).

The concept of modularity is of particular importance for evaluating information processing in the brain. Modularity was shown to be an emerging feature of networks associated with different brain functions such as learning (Bassett et al., 2011), working memory (Braun et al., 2015) and other various cognitive tasks (Bertolero et al., 2015).

Technically, a module is defined as a set of brain areas that are strongly connected to each other and weakly connected to the rest of the network. In order to quantitatively analyze the contribution of each module to the global network, two metrics, referred to as *integration* and *flexibility* can be used, as described in (Bassett et al., 2011; Braun et al., 2015). In brief, the *integration measure* describes how modules are interacting with each other while the *flexibility measure* is defined as the number of times that a node changes modular assignment normalized by the total number of possible changes (Bassett, 2011).

In the Luminous project, we plan to assess the usefulness/relevance of such measures for characterizing integrated information based on electrical brain activity. Indeed, both *integration* and *flexibility* measures can be computed on functional brain networks identified from dense-EEG data, typically during different states of consciousness, and then compared to the other metrics described in this report.

5.2.2.2 Measuring algorithmic complexity of EEG connectivity networks

Several methods are currently available to derive networks from EEG connectivity. Here we propose to study algorithmic/Kolmogorov complexity in the brain starting from such networks, assessing how compressible spatial connectivity derived networks are. These metrics will reflect algorithmic information complexity and integrated information in brain activity, and are fall in line with ideas in IIT and KT theories of consciousness. We note that algorithmic information metrics have useful properties, such as invariance under global replacement of connected and unconnected edge states. For example, Zenil et al show (Zenil et al., 2014) that small world/scale free (Barabasi-Albert and Watts-Strogatz) networks lie between regular and random networks in terms of algorithmic complexity metrics. Kolmogorov complexity approximations capture important group-theoretic and topological properties of graphs and networks, properties related to symmetry, density and connectedness. More specifically, Zenil et al. find that graphs with a large number of non-trivial automorphism groups (symmetries) tend to have smaller Kolmogorov complexity values and that graphs with differing algorithmic randomness distinguish models of networks, two of which are complex networks with different edge generation mechanisms.

The first step to derive such a metric is to compute (possibly band-passed) connectivity matrices from EEG data or EEG power (Hipp et al., 2012), preferably in source space after cortical mapping or after current source density (CSD)/laplacian estimation. Connectivity can be estimated using coherence, coherency, Granger causality (Bastos and Schoffelen, 2016), Shannon mutual information (MI), Synchronization Likelihood (SL), Mutual Algorithmic Information (MAI), etc. To study fully global algorithmic complexity, the adjacency matrices of different bands can be concatenated prior compression, or from wideband data time series (EEG or power).

Methods to avoid spurious effects on network structure from ill-posed inversion methods (Bastos and Schoffelen, 2016; Hipp et al., 2012) can be used as well. Alternatively, one can look at connectivity changes under different experimental conditions (e.g., following tCS intervention).

Adjacency matrices can then be binarized or discretized into a few levels/symbols using a threshold, and the resulting matrices (then comprised by a sequence of symbols) compressed using LZW and normalized via reshuffling (Schartner et al., 2015) or entropy estimates (Casali et al., 2013).

5.3 Machine Learning

Machine learning algorithms constitute a good approach for multi-variate analysis and therefore for the combination of individual metrics. Machine Learning may help overcoming the poor performance achieved so far in the detection of consciousness levels with more easy-to-use experimental set ups than those yielding PCI. In addition, machine learning, by combining different metrics with different flavors, may represent a practical way to account for both the integration and the differentiation of brain signals.

One of the goals within the Machine Learning strand in Luminous is to apply Deep Learning Networks, and concretely autoencoders, in order to compress the original data space. Thus, EEG data from a given cohort (e.g., MCS, UWS, anesthesia patients) can be fit through the usage of auto-encoders, the output of which will be the most informative representation of the data. The output data at the "thinnest" layer will be then compressed using LZW in order to provide an estimate of algorithmic complexity. Presumably, we will find different compressibility ratios for different consciousness levels.

Furthermore, we aim to use Machine Learning to provide a coarser discrimination among consciousness levels. So far the best discriminative metrics, i.e. PCI, achieve to establish a threshold among conscious and non-conscious states. Our final goal, which may be achieved after project ends, will be to count with a general measure able to discriminate among the different states displayed in the plot and that is able to characterize the gradual change from one to the other. This may be achieved by using a multivariate approach with the final goal of mapping a selected subset of metrics into a two-dimensional space.

6 References

Accardo, A., Affinito, M., Carrozzi, M., and Bouquet, F. (1997). Use of the fractal dimension for the analysis of electroencephalographic time series. Biol. Cybern. *77*, 339–350.

Achard, S., Delon-Martin, C., Vértes, P.E., Renard, F., Schenck, M., Schneider, F., Heinrich, C., Kremer, S., and Bullmore, E.T. (2012). Hubs of brain functional networks are radically reorganized in comatose patients. Proc. Natl. Acad. Sci. U. S. A. *109*, 20608–20613.

Akam, T., and Kullmann, D.M. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. Nat. Rev. Neurosci. *15*, 111–122.

Allegrini, P., Menicucci, D., Bedini, R., Fronzoni, L., Gemignani, A., Grigolini, P., West, B.J., and Paradisi, P. (2009). Spontaneous brain activity as a source of ideal 1/f noise. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *80*, 061914.

Ansari-Asl, K., Senhadji, L., Bellanger, J.-J., and Wendling, F. (2006). Quantitative evaluation of linear and nonlinear methods characterizing interdependencies between brain signals. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *74*, 031916.

Arthuis, M., Valton, L., Régis, J., Chauvel, P., Wendling, F., Naccache, L., Bernard, C., and Bartolomei, F. (2009). Impaired consciousness during temporal lobe seizures is related to increased long-distance cortical-subcortical synchronization. Brain J. Neurol. *132*, 2091–2101.

Aru, J., Bachmann, T., Singer, W., and Melloni, L. (2012). Distilling the neural correlates of consciousness. Neurosci. Biobehav. Rev. *36*, 737–746.

Astolfi, L., Cincotti, F., Mattia, D., Salinari, S., Babiloni, C., Basilisco, A., Rossini, P.M., Ding, L., Ni, Y., He, B., et al. (2004). Estimation of the effective and functional human cortical connectivity with structural equation modeling and directed transfer function applied to high-resolution EEG. Magn. Reson. Imaging *22*, 1457–1470.

Astolfi, L., Cincotti, F., Mattia, D., Babiloni, C., Carducci, F., Basilisco, A., Rossini, P.M., Salinari, S., Ding, L., Ni, Y., et al. (2005). Assessing cortical functional connectivity by linear inverse estimation and directed transfer function: simulations and application to real data. Clin. Neurophysiol. *116*, 920–932.

Baars, B.J. (1988). A cognitive theory of consciousness (Cambridge [England] ; New York: Cambridge University Press).

Babiloni, F., Cincotti, F., Babiloni, C., Carducci, F., Mattia, D., Astolfi, L., Basilisco, A., Rossini, P.M., Ding, L., Ni, Y., et al. (2005). Estimation of the cortical functional connectivity with the multimodal integration of high-resolution EEG and fMRI data by directed transfer function. NeuroImage *24*, 118–131.

Baccalá, L.A., and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. Biol. Cybern. *84*, 463–474.

Bachmann, T. (2000). Microgenetic Approach to the Conscious Mind (Amsterdam ; Philadelphia: John Benjamins Publishing Company).

Bai, X., Towle, V.L., He, E.J., and He, B. (2007). Evaluation of cortical current density imaging methods using intracranial electrocorticograms and functional MRI. NeuroImage *35*, 598–608.

Bandt, C., and Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. Phys. Rev. Lett. *88*, 174102.

Barttfeld, P., Uhrig, L., Sitt, J.D., Sigman, M., Jarraya, B., and Dehaene, S. (2015). Signature of consciousness in the dynamics of resting-state brain activity. Proc. Natl. Acad. Sci. U. S. A. *112*, 887–892.

Bassett, D.S., and Bullmore, E. (2006). Small-world brain networks. The Neuroscientist *12*, 512–523.

Bassett, D.S., Wymbs, N.F., Porter, M.A., Mucha, P.J., Carlson, J.M., and Grafton, S.T. (2011). Dynamic reconfiguration of human brain networks during learning. Proc. Natl. Acad. Sci. U. S. A. *108*, 7641–7646.

Bastos, A.M., and Schoffelen, J.-M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Front. Syst. Neurosci. *9*, 175.

Bédard, C., Kröger, H., and Destexhe, A. (2006). Does the 1/f frequency scaling of brain signals reflect self-organized critical states? Phys. Rev. Lett. *97*, 118102.

Beggs, J.M., and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. J. Neurosci. Off. J. Soc. Neurosci. *23*, 11167–11177.

Bekinschtein, T.A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. Proc. Natl. Acad. Sci. U. S. A. *106*, 1672–1677.

Bennett, C.H. (1988). Logical depth and physical complexity. In A Half-Century Survey on The Universal Turing Machine, (New York, NY, USA: Oxford University Press, Inc.), pp. 227–257.

Berg, P., and Scherg, M. (1994). A fast method for forward computation of multiple-shell spherical head models. Electroencephalogr. Clin. Neurophysiol. *90*, 58–64.

Bernasconi, C., and König, P. (1999). On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. Biol. Cybern. *81*, 199–210.

Bertolero, M.A., Yeo, B.T.T., and D'Esposito, M. (2015). The modular and integrative functional architecture of the human brain. Proc. Natl. Acad. Sci. U. S. A. *112*, E6798-6807.

Blumenfeld, H. (2012). Impaired consciousness in epilepsy. Lancet Neurol. *11*, 814–826.

Bola, M., and Sabel, B.A. (2015). Dynamic reorganization of brain functional networks during cognition. NeuroImage *114*, 398–413.

Boly, M., Moran, R., Murphy, M., Boveroux, P., Bruno, M.-A., Noirhomme, Q., Ledoux, D., Bonhomme, V., Brichant, J.-F., Tononi, G., et al. (2012). Connectivity changes underlying spectral EEG changes during propofol-induced loss of consciousness. J. Neurosci. *32*, 7082–7090.

Bonini, F., Lambert, I., Wendling, F., McGonigal, A., and Bartolomei, F. (2016). Altered synchrony and loss of consciousness during frontal lobe seizures. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. *127*, 1170–1175.

Braun, U., Schäfer, A., Walter, H., Erk, S., Romanczuk-Seiferth, N., Haddad, L., Schweiger, J.I., Grimm, O., Heinz, A., Tost, H., et al. (2015). Dynamic reconfiguration of frontal brain networks during executive cognition in humans. Proc. Natl. Acad. Sci. U. S. A. *112*, 11678–11683.

Brazier, M.A. (1968). Electrical activity recorded simultaneously from the scalp and deep structures of the human brain. A computer study of their relationships. J. Nerv. Ment. Dis. *147*, 31–39.

Brazier, M. a. B., and Casby, J.U. (1952). Cross-correlation and autocorrelation studies of electroencephalographic potentials. Electroencephalogr. Clin. Neurophysiol. *4*, 201–211.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Brenner, R.P. (2005). The interpretation of the EEG in stupor and coma. The Neurologist *11*, 271–284.

Brookes, M.J., Hale, J.R., Zumer, J.M., Stevenson, C.M., Francis, S.T., Barnes, G.R., Owen, J.P., Morris, P.G., and Nagarajan, S.S. (2011). Measuring functional connectivity using MEG: methodology and comparison with fcMRI. NeuroImage *56*, 1082–1104.

Brown, E.N., Lydic, R., and Schiff, N.D. (2010). General anesthesia, sleep, and coma. N. Engl. J. Med. *363*, 2638–2650.

Buiatti, M., Papo, D., Baudonnière, P.-M., and van Vreeswijk, C. (2007). Feedback modulates the temporal scale-free dynamics of brain electrical activity in a hypothesis testing task. Neuroscience *146*, 1400–1412.

Butts, C.T. (2001). The complexity of social networks: theoretical and empirical findings. Soc. Netw. *23*, 31–72.

Caparros-Lefebvre, D., Destee, A., and Petit, H. (1997). Late onset familial dystonia: could mitochondrial deficits induce a diffuse lesioning process of the whole basal ganglia system? J. Neurol. Neurosurg. Psychiatry *63*, 196–203.

Casali, A.G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K.R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. Sci. Transl. Med. *5*, 198ra105.

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., G Casali, A., Trimarchi, P.D., Boly, M., et al. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. Ann. Neurol. *80*, 718-729.

Chaitin, G.J. (1993). Randomness in arithmetic and the decline and fall of reductionism in pure mathematics. ArXivchao-Dyn9304002.

Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibáñez, A., Owen, A.M., and Bekinschtein, T.A. (2013). Expectation and attention in hierarchical auditory prediction. J. Neurosci. *33*, 11194–11205.

Chennu, S., Finoia, P., Kamau, E., Allanson, J., Williams, G.B., Monti, M.M., Noreika, V., Arnatkeviciute, A., Canales-Johnson, A., Olivares, F., et al. (2014). Spectral

signatures of reorganised brain networks in disorders of consciousness. PLoS Comput. Biol. *10*, e1003887.

Cimenser, A., Purdon, P.L., Pierce, E.T., Walsh, J.L., Salazar-Gomez, A.F., Harrell, P.G., Tavares-Stoeckel, C., Habeeb, K., and Brown, E.N. (2011). Tracking brain states under general anesthesia by using global coherence analysis. Proc. Natl. Acad. Sci. U. S. A. *108*, 8832–8837.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. *36*, 181–204.

Cote, K.A. (2002). Probing awareness during sleep with the auditory odd-ball paradigm. Int. J. Psychophysiol. *46*, 227–241.

Cover, T.M., and Thomas, J.A. (1991). Elements of Information Theory 2nd Edition (Hoboken, N.J: Wiley-Interscience).

Crick, F., and Koch, C. (1990). Towards a neurobiological theory of consciousness. In Seminars in the Neurosciences (Saunders Scientific Publications), pp. 263–275.

Dale, A.M., and Sereno, M.I. (1993). Improved localizadon of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. J. Cogn. Neurosci. *5*, 162–176.

David, O., Garnero, L., Cosmelli, D., and Varela, F.J. (2002). Estimation of neural dynamics from MEG/EEG cortical current density maps: application to the reconstruction of large-scale cortical synchrony. IEEE Trans. Biomed. Eng. *49*, 975–987.

David, O., Cosmelli, D., Hasboun, D., and Garnero, L. (2003). A multitrial analysis for revealing significant corticocortical networks in magnetoencephalography and electroencephalography. NeuroImage *20*, 186–201.

David, O., Cosmelli, D., and Friston, K.J. (2004). Evaluation of different measures of functional connectivity using a neural mass model. NeuroImage *21*, 659–673.

Davis, P.A. (1939). Effects of acoustic stimuli on the waking human brain. J. Neurophysiol. *2*, 494–499.

De Los Rios, P., and Zhang, Y.-C. (1999). Universal 1/f Noise from Dissipative Self-Organized Criticality Models. Phys. Rev. Lett. *82*, 472–475.

Dehaene, S., and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. Neuron *70*, 200–227.

Dehaene, S., Kerszberg, M., and Changeux, J.P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. Proc. Natl. Acad. Sci. U. S. A. *95*, 14529–14534.

Dehaene, S., Sergent, C., and Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc. Natl. Acad. Sci. U. S. A. *100*, 8520–8525.

Dehaene, S., Charles, L., King, J.-R., and Marti, S. (2014). Toward a computational theory of conscious processing. Curr. Opin. Neurobiol. *25*, 76–84.

Descartes, R. (1999). Discourse on Method and Meditations on First Philosophy, 4th Ed. (Indianapolis: Hackett Publishing Company).

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. Data Min. Knowl. Discov. *3*, 409–425.

Drakesmith, M., El-Deredy, W., and Welbourne, S. (2013). Reconstructing coherent networks from electroencephalography and magnetoencephalography with reduced contamination from volume conduction or magnetic field spread. PloS One *8*, e81553.

Dubovikov, M.., Starchenko, N.., and Dubovikov, M.. (2004). Dimension of the minimal cover and fractal analysis of time series. Phys. Stat. Mech. Its Appl. *339*, 591–608.

Duncan-Johnson, C.C., and Donchin, E. (1977). On quantifying surprise: the variation of event-related potentials with subjective probability. Psychophysiology *14*, 456–467.

Eddy, M., Schmid, A., and Holcomb, P.J. (2006). Masked repetition priming and event-related brain potentials: a new approach for tracking the time-course of object perception. Psychophysiology *43*, 564–568.

Eguíluz, V.M., Chialvo, D.R., Cecchi, G.A., Baliki, M., and Apkarian, A.V. (2005). Scale-free brain functional networks. Phys. Rev. Lett. *94*, 018102.

Faugeras, F., Rohaut, B., Weiss, N., Bekinschtein, T., Galanaud, D., Puybasset, L., Bolgert, F., Sergent, C., Cohen, L., Dehaene, S., et al. (2012). Event related potentials elicited by violations of auditory regularities in patients with impaired consciousness. Neuropsychologia *50*, 403–418.

Fernández-Espejo, D., and Owen, A.M. (2013). Detecting awareness after severe brain injury. Nat. Rev. Neurosci. *14*, 801–809.

Fernández-Espejo, D., Bekinschtein, T., Monti, M.M., Pickard, J.D., Junque, C., Coleman, M.R., and Owen, A.M. (2011). Diffusion weighted imaging distinguishes the vegetative state from the minimally conscious state. NeuroImage *54*, 103–112.

Fernández-Espejo, D., Rossit, S., and Owen, A.M. (2015). A thalamocortical mechanism for the absence of overt motor behavior in covertly aware patients. JAMA Neurol. *72*, 1442–1450.

Fischer, C., Luauté, J., Adeleine, P., and Morlet, D. (2004). Predictive value of sensory and cognitive evoked potentials for awakening from coma. Neurology *63*, 669–673.

Fischer, C., Luaute, J., and Morlet, D. (2010). Event-related potentials (MMN and novelty P3) in permanent vegetative or minimally conscious states. Clin. Neurophysiol. *121*, 1032–1042.

Forgacs, P.B., Conte, M.M., Fridman, E.A., Voss, H.U., Victor, J.D., and Schiff, N.D. (2014). Preservation of electroencephalographic organization in patients with impaired

consciousness and imaging-based evidence of command-following. Ann. Neurol. *76*, 869–879.

Fries, P. (2015). Rhythms for cognition: communication through coherence. Neuron *88*, 220–235.

Friston, K., Henson, R., Phillips, C., and Mattout, J. (2006). Bayesian estimation of evoked and induced responses. Hum. Brain Mapp. *27*, 722–735.

Friston, K., Moran, R., and Seth, A.K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling. Curr. Opin. Neurobiol. *23*, 172–178.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. J. Am. Stat. Assoc. *77*, 304–313.

Ghuman, A.S., McDaniel, J.R., and Martin, A. (2011). A wavelet-based method for measuring the oscillatory dynamics of resting-state functional connectivity in MEG. NeuroImage *56*, 69–77.

Giacino, J.T., Kalmar, K., and Whyte, J. (2004). The JFK Coma Recovery Scale-Revised: Measurement characteristics and diagnostic utility. Arch. Phys. Med. Rehabil. *85*, 2020–2029.

Godwin, D., Barry, R.L., and Marois, R. (2015). Breakdown of the brain's functional network modularity with awareness. Proc. Natl. Acad. Sci. *112*, 3799–3804.

Gökyiğit, A., and Calişkan, A. (1995). Diffuse spike-wave status of 9-year duration without behavioral change or intellectual decline. Epilepsia *36*, 210–213.

Gómez, C., Mediavilla, A., Hornero, R., Abásolo, D., and Fernández, A. (2009). Use of the Higuchi's fractal dimension for the analysis of MEG recordings from Alzheimer's disease patients. Med. Eng. Phys. *31*, 306–313.

Gómez, C.M., Delinte, A., Vaquero, E., Cardoso, M.J., Vázquez, M., Crommelinck, M., and Roucoux, A. (2001). Current source density analysis of CNV during temporal gap paradigm. Brain Topogr. *13*, 149–159.

Gosseries, O., Schnakers, C., Ledoux, D., Vanhaudenhuyse, A., Bruno, M.-A., Demertzi, A., Noirhomme, Q., Lehembre, R., Damas, P., Goldman, S., et al. (2011). Automated EEG entropy measurements in coma, vegetative state/unresponsive wakefulness syndrome and minimally conscious state. Funct. Neurol. *26*, 25–30.

Granger, C.W. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica 424–438.

Gray, C.M., König, P., Engel, A.K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature *338*, 334–337.

Grova, C., Daunizeau, J., Lina, J.-M., Bénar, C.G., Benali, H., and Gotman, J. (2006). Evaluation of EEG localization methods using realistic simulations of interictal spikes. NeuroImage *29*, 734–753.

Grunwald, P., and Vitanyi, P. (2004). Shannon Information and Kolmogorov Complexity. ArXivcs0410002 CsIT.

Güler, I., and Ubeyli, E.D. (2007). Multiclass support vector machines for EEG-signals classification. IEEE Trans. Inf. Technol. Biomed. Publ. IEEE Eng. Med. Biol. Soc. *11*, 117–126.

Hallett, M. (2000). Transcranial magnetic stimulation and the human brain. Nature *406*, 147–150.

Hämäläinen, M.S., and Ilmoniemi, R.J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. Med. Biol. Eng. Comput. *32*, 35–42.

Hamilton, J.D. (1994). Time Series Analysis (Princeton, N.J: Princeton University Press).

Harrison, A.H., and Connolly, J.F. (2013). Finding a way in: A review and practical evaluation of fMRI and EEG for detection and assessment in disorders of consciousness. Neurosci. Biobehav. Rev. *37*, 1403–1419.

Hassan, M., Dufor, O., Merlet, I., Berrou, C., and Wendling, F. (2014). EEG source connectivity analysis: from dense array recordings to brain networks. PloS One *9*, e105041.

He, B.J., Zempel, J.M., Snyder, A.Z., and Raichle, M.E. (2010). The temporal structures and functional significance of scale-free brain activity. Neuron *66*, 353–369.

Helfrich, R.F., Schneider, T.R., Rach, S., Trautmann-Lengsfeld, S.A., Engel, A.K., and Herrmann, C.S. (2014). Entrainment of brain oscillations by transcranial alternating current stimulation. Curr. Biol. *24*, 333–339.

Helfrich, R.F., Knepper, H., Nolte, G., Sengelmann, M., König, P., Schneider, T.R., and Engel, A.K. (2016). Spectral fingerprints of large-scale cortical dynamics during ambiguous motion perception. Hum. Brain Mapp. *37*, 4099–4111.

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. Proc. Natl. Acad. Sci. U. S. A. *109 Suppl 1*, 10661–10668.

Herculano-Houzel, S., Munk, M.H., Neuenschwander, S., and Singer, W. (1999). Precisely synchronized oscillatory firing patterns require electroencephalographic activation. J. Neurosci. *19*, 3992–4010.

Hermes, D., Miller, K.J., Wandell, B.A., and Winawer, J. (2015). Stimulus dependence of gamma oscillations in human visual cortex. Cereb. Cortex *25*, 2951–2959.

Hesse, J., and Gross, T. (2014). Self-organized criticality as a fundamental property of neural systems. Front. Syst. Neurosci. *8*, 166.

van den Heuvel, M.P., and Sporns, O. (2011). Rich-club organization of the human connectome. J. Neurosci. *31*, 15775–15786.

Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. Phys. Nonlinear Phenom. *31*, 277–283.

Hipp, J.F., Hawellek, D.J., Corbetta, M., Siegel, M., and Engel, A.K. (2012). Large-scale cortical correlation structure of spontaneous oscillatory activity. Nat. Neurosci. *15*, 884–890.

Hoechstetter, K., Bornfleth, H., Weckesser, D., Ille, N., Berg, P., and Scherg, M. (2004). BESA source coherence: a new method to study cortical oscillatory coupling. Brain Topogr. *16*, 233–238.

Hoel, E.P., Albantakis, L., and Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. Proc. Natl. Acad. Sci. U. S. A. *110*, 19790–19795.

Hohwy, J. (2013). The Predictive Mind (Oxford University Press).

Holcombe, A.O. (2009). Seeing slow and seeing fast: two limits on perception. Trends Cogn. Sci. *13*, 216–221.

Höller, Y., Bergmann, J., Kronbichler, M., Crone, J.S., Schmid, E.V., Golaszewski, S., and Ladurner, G. (2011). Preserved oscillatory response but lack of mismatch negativity in patients with disorders of consciousness. Clin. Neurophysiol. *122*, 1744–1754.

Höller, Y., Thomschewski, A., Bergmann, J., Kronbichler, M., Crone, J.S., Schmid, E.V., Butz, K., Höller, P., Nardone, R., and Trinka, E. (2014). Connectivity biomarkers can differentiate patients with different levels of consciousness. Clin. Neurophysiol. *125*, 1545–1555.

Horki, P., Bauernfeind, G., Klobassa, D.S., Pokorny, C., Pichler, G., Schippinger, W., and Müller-Putz, G.R. (2014). Detection of mental imagery and attempted movements in patients with disorders of consciousness using EEG. Front. Hum. Neurosci. *8*, 1009.

Hudetz, A.G., Liu, X., and Pillay, S. (2015). Dynamic repertoire of intrinsic brain states is reduced in propofol-induced unconsciousness. Brain Connect. *5*, 10–22.

Hutter, M. (2007). Algorithmic information theory. Scholarpedia *2*, 2519.

Ihlen, E.A.F. (2012). Introduction to multifractal detrended fluctuation analysis in matlab. Front. Physiol. *3*, 141.

Imas, O.A., Ropella, K.M., Ward, B.D., Wood, J.D., and Hudetz, A.G. (2005). Volatile anesthetics disrupt frontal-posterior recurrent information transfer at gamma frequencies in rat. Neurosci. Lett. *387*, 145–150.

Jaeger, H. (2001). The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. German National Research Center for Information Technology GMD Technical Report *148*, 34.

Jain, S.K., Sundar, I.V., Sharma, V., Prasanna, K.L., Kulwal, G., and Tiwari, R.N. (2013). Bilateral large traumatic basal ganglia haemorrhage in a conscious adult: a rare case report. Brain Inj. *27*, 500–503.

Jiang, J. (1999). Image compression with neural networks - A survey. In Signal Processing: Image Communication, pp. 737–760.

Jordan, D., Stockmanns, G., Kochs, E.F., Pilge, S., and Schneider, G. (2008). Electroencephalographic order pattern analysis for the separation of consciousness and unconsciousness: an analysis of approximate entropy, permutation entropy, recurrence rate, and phase coupling of order recurrence plots. Anesthesiology *109*, 1014–1022.

Kalauzi, A., Bojić, T., and Rakić, L. (2009). Extracting complexity waveforms from one-dimensional signals. Nonlinear Biomed. Phys. *3*, 8.

Kamiński, M.J., and Blinowska, K.J. (1991). A new method of the description of the information flow in the brain structures. Biol. Cybern. *65*, 203–210.

Kaskinoro, K., Maksimow, A., Långsjö, J., Aantaa, R., Jääskeläinen, S., Kaisti, K., Särkelä, M., and Scheinin, H. (2011). Wide inter-individual variability of bispectral index and spectral entropy at loss of consciousness during increasing concentrations of dexmedetomidine, propofol, and sevoflurane. Br. J. Anaesth. *107*, 573–580.

Kaspar, F., and Schuster, H. (1987). Easily calculable measure for the complexity of spatiotemporal patterns. Phys. Rev. A *36*, 842–848.

Kello, C.T., Brown, G.D.A., Ferrer-I-Cancho, R., Holden, J.G., Linkenkaer-Hansen, K., Rhodes, T., and Van Orden, G.C. (2010). Scaling laws in cognitive sciences. Trends Cogn. Sci. *14*, 223–232.

King, J.R., Faugeras, F., Gramfort, A., Schurger, A., El Karoui, I., Sitt, J.D., Rohaut, B., Wacongne, C., Labyt, E., Bekinschtein, T., et al. (2013a). Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. NeuroImage *83*, 726–738.

King, J.-R., Sitt, J.D., Faugeras, F., Rohaut, B., El Karoui, I., Cohen, L., Naccache, L., and Dehaene, S. (2013b). Information sharing in the brain indexes consciousness in noncommunicative patients. Curr. Biol. *23*, 1914–1919.

Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. Nat. Rev. Neurosci. *17*, 307–321.

Kotchoubey, B. (2005). Event-related potential measures of consciousness: two equations with three unknowns. Prog. Brain Res. *150*, 427–444.

Lambert, I., Arthuis, M., McGonigal, A., Wendling, F., and Bartolomei, F. (2012). Alteration of global workspace during loss of consciousness: a study of parietal seizures. Epilepsia *53*, 2104–2110.

Latora, V., and Marchiori, M. (2001). Efficient behavior of small-world networks. Phys. Rev. Lett. *87*, 198701.

Laureys, S., and Schiff, N.D. (2012). Coma and consciousness: Paradigms (re)framed by neuroimaging. NeuroImage *61*, 478–491.

Lee, A.T., Gee, S.M., Vogt, D., Patel, T., Rubenstein, J.L., and Sohal, V.S. (2014). Pyramidal neurons in prefrontal cortex receive subtype-specific forms of excitation and inhibition. Neuron *81*, 61–68.

Lee, J., Kim, D., and Shin, H.-S. (2004). Lack of delta waves and sleep disturbances during non-rapid eye movement sleep in mice lacking alpha1G-subunit of T-type calcium channels. Proc. Natl. Acad. Sci. U. S. A. *101*, 18195–18199.

Lee, J.M., Kim, D.J., Kim, I.Y., Park, K.S., and Kim, S.I. (2002). Detrended fluctuation analysis of EEG in sleep apnea using MIT/BIH polysomnography data. Comput. Biol. Med. *32*, 37–47.

Leistedt, S., Dumont, M., Lanquart, J.-P., Jurysta, F., and Linkowski, P. (2007). Characterization of the sleep EEG in acutely depressed men using detrended fluctuation analysis. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. *118*, 940–950.

Lemon, R.N., and Edgley, S.A. (2010). Life without a cerebellum. Brain *133*, 652–654.

Lempel, A., and Ziv, J. (1976). On the complexity of finite sequences. IEEE Trans. Inf. Theory *22*, 75–81.

Li, M., and Vitanyi, P. (1997). An Introduction to Kolmogorov Complexity and Its Applications (Springer-Verlag New York).

Li, X., Cui, S., and Voss, L.J. (2008). Using permutation entropy to measure the electroencephalographic effects of sevoflurane. Anesthesiology *109*, 448–456.

Lima, B., Singer, W., and Neuenschwander, S. (2011). Gamma responses correlate with temporal expectation in monkey primary visual cortex. J. Neurosci. *31*, 15919–15931.

Lin, F.-H., Witzel, T., Hämäläinen, M.S., Dale, A.M., Belliveau, J.W., and Stufflebeam, S.M. (2004). Spectral spatiotemporal imaging of cortical oscillations and interactions in the human brain. NeuroImage *23*, 582–595.

Lopes, R., and Betrouni, N. (2009). Fractal and multifractal analysis: a review. Med. Image Anal. *13*, 634–649.

Ludwig, W., and Falter, C. (2012). Symmetries in Physics: Group Theory Applied to Physical Problems (Springer Science & Business Media).

Luo, Q., Mitchell, D., Cheng, X., Mondillo, K., Mccaffrey, D., Holroyd, T., Carver, F., Coppola, R., and Blair, J. (2009). Visual awareness, emotion, and gamma band synchronization. Cereb. Cortex *19*, 1896–1904.

Lv, J., Simpson, D.M., and Bell, S.L. (2007). Objective detection of evoked potentials using a bootstrap technique. Med. Eng. Phys. *29*, 191–198.

Lyytinen, H., Blomberg, A.P., and Näätänen, R. (1992). Event-related potentials and autonomic responses to a change in unattended auditory stimuli. Psychophysiology *29*, 523–534.

Marinazzo, D., Gosseries, O., Boly, M., Ledoux, D., Rosanova, M., Massimini, M., Noirhomme, Q., and Laureys, S. (2014). Directed information transfer in scalp electroencephalographic recordings: insights on disorders of consciousness. Clin. EEG Neurosci. *45*, 33–39.

Mars, N.J., and Lopes da Silva, F.H. (1983). Propagation of seizure activity in kindled dogs. Electroencephalogr. Clin. Neurophysiol. *56*, 194–209.

Mattout, J., Phillips, C., Penny, W.D., Rugg, M.D., and Friston, K.J. (2006). MEG source localization under multiple constraints: an extended Bayesian framework. NeuroImage *30*, 753–767.

McCormick, D.A., Wang, Z., and Huguenard, J. (1993). Neurotransmitter control of neocortical neuronal activity and excitability. Cereb. Cortex *3*, 387–398.

Meeren, H.K.M., Pijn, J.P.M., Van Luijtelaar, E.L.J.M., Coenen, A.M.L., and Lopes da Silva, F.H. (2002). Cortical focus drives widespread corticothalamic networks during spontaneous absence seizures in rats. J. Neurosci. Off. J. Soc. Neurosci. *22*, 1480–1495.

Meisel, C., Olbrich, E., Shriki, O., and Achermann, P. (2013). Fading signatures of critical brain dynamics during sustained wakefulness in humans. J. Neurosci. Off. J. Soc. Neurosci. *33*, 17363–17372.

Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., and Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. J. Neurosci. *27*, 2858–2865.

Monti, M.M., Lutkenhoff, E.S., Rubinov, M., Boveroux, P., Vanhaudenhuyse, A., Gosseries, O., Bruno, M.-A., Noirhomme, Q., Boly, M., and Laureys, S. (2013). Dynamic change of global and local information processing in propofol-induced loss and recovery of consciousness. PLoS Comput. Biol. *9*, e1003271.

Morlet, D., and Fischer, C. (2014). MMN and novelty P3 in coma and other altered states of consciousness: a review. Brain Topogr. *27*, 467–479.

Moruzzi, G., and Magoun, H.W. (1949). Brain stem reticular formation and activation of the EEG. Electroencephalogr. Clin. Neurophysiol. *1*, 455–473.

Munk, M.H., Roelfsema, P.R., König, P., Engel, A.K., and Singer, W. (1996). Role of reticular activation in the modulation of intracortical synchronization. Science *272*, 271–274.

Murphy, M., Bruno, M.-A., Riedner, B.A., Boveroux, P., Noirhomme, Q., Landsness, E.C., Brichant, J.-F., Phillips, C., Massimini, M., Laureys, S., et al. (2011). Propofol anesthesia and sleep: a high-density EEG study. Sleep *34*, 283–291A.

Näätänen, R., and Picton, T.W. (1986). N2 and automatic versus controlled processes. Electroencephalogr. Clin. Neurophysiol. Suppl. *38*, 169–186.

Näätänen, R., and Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. Psychol. Bull. *125*, 826–859.

Näätänen, R., Gaillard, A.W., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. (Amst.) *42*, 313–329.

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clin. Neurophysiol. *118*, 2544–2590.

Naccache, L., Puybasset, L., Gaillard, R., Serve, E., and Willer, J.-C. (2005). Auditory mismatch negativity is a good predictor of awakening in comatose patients: a fast and reliable procedure. Clin. Neurophysiol. *116*, 988–989.

Nagai, Y., Critchley, H.D., Featherstone, E., Fenwick, P.B.C., Trimble, M.R., and Dolan, R.J. (2004). Brain activity relating to the contingent negative variation: an fMRI investigation. NeuroImage *21*, 1232–1241.

Nashida, T., Yabe, H., Sato, Y., Hiruma, T., Sutoh, T., Shinozaki, N., and Kaneko, S. (2000). Automatic auditory information processing in sleep. Sleep *23*, 821–828.

Nobili, L., De Gennaro, L., Proserpio, P., Moroni, F., Sarasso, S., Pigorini, A., De Carli, F., and Ferrara, M. (2012). Local aspects of sleep: observations from intracerebral recordings in humans. Prog. Brain Res. *199*, 219–232.

Noirhomme, Q., Brecheisen, R., Lesenfants, D., Antonopoulos, G., and Laureys, S. (2015). "Look at my classifier's result": Disentangling unresponsive from (minimally) conscious patients. NeuroImage.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., and Hallett, M. (2004). Identifying true brain interaction from EEG data using the imaginary part of coherency. Clin. Neurophysiol. *115*, 2292–2307.

Nunez, P.L., and Srinivasan, R. (2006). A theoretical basis for standing and traveling brain waves measured with human EEG with implications for an integrated consciousness. Clin. Neurophysiol. *117*, 2424–2435.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput. Biol. *10*, e1003588.

Olofsen, E., Sleigh, J.W., and Dahan, A. (2008). Permutation entropy of the electroencephalogram: a measure of anaesthetic drug effect. Br. J. Anaesth. *101*, 810–821.

Paavilainen, P., Simola, J., Jaramillo, M., Näätänen, R., and Winkler, I. (2001). Preattentive extraction of abstract feature conjunctions from auditory stimulation as reflected by the mismatch negativity (MMN). Psychophysiology *38*, 359–365.

Palva, J.M., Palva, S., and Kaila, K. (2005). Phase Synchrony among neuronal oscillations in the human cortex. J. Neurosci. *25*, 3962–3972.

Pantazis, D., Nichols, T.E., Baillet, S., and Leahy, R.M. (2005). A comparison of random field theory and permutation methods for the statistical analysis of MEG data. NeuroImage *25*, 383–394.

Paradisi, P., Allegrini, P., Gemignani, A., Laurino, M., Menicucci, D., and Piarulli, A. (2013). Scaling and intermittency of brain events as a manifestation of consciousness. In AIP Conference Proceedings, pp. 151–161.

Pascual-Marqui, R.D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. Methods Find. Exp. Clin. Pharmacol. *24 Suppl D*, 5–12.

Pascual-Marqui, R.D., Michel, C.M., and Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. Int. J. Psychophysiol. *18*, 49–65.

Passerini, F., and Severini, S. (2008). The von Neumann entropy of networks. Available SSRN 1382662.

Pesaran, B., Pezaris, J.S., Sahani, M., Mitra, P.P., and Andersen, R.A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nat. Neurosci. *5*, 805–811.

Peterson, A., Cruse, D., Naci, L., Weijer, C., and Owen, A.M. (2015). Risk, diagnostic error, and the clinical science of consciousness. NeuroImage Clin. *7*, 588–597.

Phillips, C., Mattout, J., Rugg, M.D., Maquet, P., and Friston, K.J. (2005). An empirical Bayesian solution to the source reconstruction problem in EEG. NeuroImage *24*, 997–1011.

Picton, T.W. (1992). The P300 wave of the human event-related potential. J. Clin. Neurophysiol. *9*, 456–479.

Pigorini, A., Sarasso, S., Proserpio, P., Szymanski, C., Arnulfo, G., Casarotto, S., Fecchio, M., Rosanova, M., Mariotti, M., Lo Russo, G., et al. (2015). Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep. NeuroImage *112*, 105–113.

Pijn, J.P., and Silva, F.L. da (1993). Propagation of Electrical Activity: Nonlinear Associations and Time Delays between EEG Signals. In Basic Mechanisms of the EEG, S. Zschocke, and E.-J. Speckmann, eds. (Birkhäuser Boston), pp. 41–61.

Pijn, J.P., Vijn, P.C., Lopes da Silva, F.H., Van Ende Boas, W., and Blanes, W. (1990). Localization of epileptogenic foci using a new signal analytical approach. Neurophysiol. Clin. Clin. Neurophysiol. *20*, 1–11.

Pipa, G., Städtler, E.S., Rodriguez, E.F., Waltz, J.A., Muckli, L.F., Singer, W., Goebel, R., and Munk, M.H.J. (2009). Performance- and stimulus-dependent oscillations in monkey prefrontal cortex during short-term memory. Front. Integr. Neurosci. *3*, 25.

Pitts, M.A., Martínez, A., and Hillyard, S.A. (2012). Visual processing of contour patterns under conditions of inattentional blindness. J. Cogn. Neurosci. *24*, 287–303.

Pitts, M.A., Metzler, S., and Hillyard, S.A. (2014). Isolating neural correlates of conscious perception from neural correlates of reporting one's perception. Front. Psychol. *5*, 1078.

Pockett, S., and Holmes, M.D. (2009). Intracranial EEG power spectra and phase synchrony during consciousness and unconsciousness. Conscious. Cogn. *18*, 1049–1055.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. *118*, 2128–2148.

Pöppel, E. (1988). Size constancy and oculomotor modulation of perifoveal light-difference threshold. Naturwissenschaften *75*, 463–465.

Posner, M.I., Klein, R., Summers, J., and Buggie, S. (1973). On the selection of signals. Mem. Cognit. *1*, 2–12.

Priesemann, V., Valderrama, M., Wibral, M., and Le Van Quyen, M. (2013). Neuronal avalanches differ from wakefulness to deep sleep--evidence from intracranial depth recordings in humans. PLoS Comput. Biol. *9*, e1002985.

Prokopenko, M., Boschetti, F., and Ryan, A.J. (2009). An information-theoretic primer on complexity, self-organization, and emergence. Complexity *15*, 11–28.

Ray, A., Tao, J.X., Hawes-Ebersole, S.M., and Ebersole, J.S. (2007). Localizing value of scalp EEG spikes: a simultaneous scalp and intracranial study. Clin. Neurophysiol. *118*, 69–79.

Rodriguez, E., George, N., Lachaux, J.P., Martinerie, J., Renault, B., and Varela, F.J. (1999). Perception's shadow: long-distance synchronization of human brain activity. Nature *397*, 430–433.

Roelfsema, P.R., Engel, A.K., König, P., and Singer, W. (1997). Visuomotor integration is associated with zero time-lag synchronization among cortical areas. Nature *385*, 157–161.

Rosenblum, M.G., and Pikovsky, A.S. (2001). Detecting direction of coupling in interacting oscillators. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *64*, 045202.

Ruffini, G. (2007). Information, complexity, brains and reality (Kolmogorov Manifesto). ArXiv07041147 Physicsgen-Ph.

Ruffini, G. (2009). Reality as Simplicity. ArXiv09031193 Physicsgen-Ph.

Ruffini, G. (2015). Application of the reciprocity theorem to EEG inversion and optimization of EEG-driven transcranial current stimulation (tCS, including tDCS, tACS, tRNS). ArXiv150604835 Physicsbio-Ph.

Ruffini, G. (2016). An Algorithmic Information Theory of Consciousness (KT), Starlab Technical Note TN00338, submitted to Neuroscience of Consciousness.

Rutiku, R., Martin, M., Bachmann, T., and Aru, J. (2015). Does the P300 reflect conscious perception or its consequences? Neuroscience *298*, 180–189.

Saarinen, J., Paavilainen, P., Schöger, E., Tervaniemi, M., and Näätänen, R. (1992). Representation of abstract attributes of auditory stimuli in the human brain. Neuroreport *3*, 1149–1151.

Sanders, R.D., Tononi, G., Laureys, S., and Sleigh, J. (2012). Unresponsiveness ≠ Unconsciousness. Anesthesiology *116*, 946–959.

Sanei, S., and Chambers, J.A. (2008). EEG Signal Processing (Wiley).

Sarà, M., Pistoia, F., Pasqualetti, P., Sebastiano, F., Onorati, P., and Rossini, P.M. (2011). Functional isolation within the cerebral cortex in the vegetative state: A nonlinear method to predict clinical outcomes. Neurorehabil. Neural Repair *25*, 35–42.

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, A.G., Brichant, J.-F., Boveroux, P., et al. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. Curr. Biol. *25*, 3099–3105.

Schartner, M., Seth, A., Noirhomme, Q., Boly, M., Bruno, M.-A., Laureys, S., and Barrett, A. (2015). Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia. PloS One *10*, e0133532.

Schiff, N.D. (2009). Central thalamic deep-brain stimulation in the severely injured brain: rationale and proposed mechanisms of action. Ann. N. Y. Acad. Sci. *1157*, 101–116.

Schiff, N.D. (2010). Recovery of consciousness after brain injury: a mesocircuit hypothesis. Trends Neurosci. *33*, 1–9.

Schnakers, C., Vanhaudenhuyse, A., Giacino, J., Ventura, M., Boly, M., Majerus, S., Moonen, G., and Laureys, S. (2009). Diagnostic accuracy of the vegetative and minimally conscious state: clinical consensus versus standardized neurobehavioral assessment. BMC Neurol. *9*, 35.

Schneider, W., and Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. Psychol. Rev. *84*, 1–66.

Schoffelen, J.-M., and Gross, J. (2009). Source connectivity analysis with MEG and EEG. Hum. Brain Mapp. *30*, 1857–1865.

Sergent, C., Baillet, S., and Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. Nat. Neurosci. *8*, 1391–1400.

Seth, A.K. (2005). Causal connectivity of evolved neural networks during behavior. Network *16*, 35–54.

Seth, A.K. (2013). Interoceptive inference, emotion, and the embodied self. Trends Cogn. Sci. *17*, 565–573.

Seth, A.K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. Cogn. Neurosci. *5*, 97–118.

Seth, A.K., and Edelman, G.M. (2004). Environment and behavior influence the complexity of evolved neural networks. Adapt. Behav. *12*, 5–20.

Shanahan, M. (2010). Metastable chimera states in community-structured oscillator networks. Chaos *20*, 013108.

Sigl, J.C., and Chamoun, N.G. (1994). An introduction to bispectral analysis for the electroencephalogram. J. Clin. Monit. *10*, 392–404.

Silverstein, B.H., Snodgrass, M., Shevrin, H., and Kushwaha, R. (2015). P3b, consciousness, and complex unconscious processing. Cortex *73*, 216–227.

Singer, W. (1999). Time as coding space? Curr. Opin. Neurobiol. *9*, 189–194.

Singer, W., and Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. Annu. Rev. Neurosci. *18*, 555–586.

Sitt, J.D., King, J.-R., El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S., and Naccache, L. (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. Brain *137*, 2258–2270.

Solovey, G., Alonso, L.M., Yanagawa, T., Fujii, N., Magnasco, M.O., Cecchi, G.A., and Proekt, A. (2015). Loss of consciousness is associated with stabilization of cortical activity. J. Neurosci. *35*, 10866–10877.

Sporns, O. (2011). The human connectome: a complex network. Ann. N. Y. Acad. Sci. *1224*, 109–125.

Sporns, O., and Betzel, R.F. (2016). Modular brain networks. Annu. Rev. Psychol. *67*, 613–640.

Sporns, O., and Tononi, G. (2002). Classes of network connectivity and dynamics. Complexity *7*, 28–38.

Sporns, O., Tononi, G., and Edelman, G.M. (2000). Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. Cereb. Cortex *10*, 127–141.

Squires, N.K., Squires, K.C., and Hillyard, S.A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. Electroencephalogr. Clin. Neurophysiol. *38*, 387–401.

Stam, C.J., Breakspear, M., van Cappellen van Walsum, A.-M., and van Dijk, B.W. (2003). Nonlinear synchronization in EEG and whole-head MEG recordings of healthy subjects. Hum. Brain Mapp. *19*, 63–78.

Steriade, M. (2000). Corticothalamic resonance, states of vigilance and mentation. Neuroscience *101*, 243–276.

Steriade, M., Amzica, F., and Contreras, D. (1996). Synchronization of fast (30-40 Hz) spontaneous cortical rhythms during brain activation. J. Neurosci. *16*, 392–417.

Steriade, M., Timofeev, I., and Grenier, F. (2001). Natural waking and sleep states: a view from inside neocortical neurons. J. Neurophysiol. *85*, 1969–1985.

Straussberg, R., Shorer, Z., Weitz, R., Basel, L., Kornreich, L., Corie, C.I., Harel, L., Djaldetti, R., and Amir, J. (2002). Familial infantile bilateral striatal necrosis: clinical features and response to biotin treatment. Neurology *59*, 983–989.

Supp, G.G., Schlögl, A., Trujillo-Barreto, N., Müller, M.M., and Gruber, T. (2007). Directed cortical information flow during human object recognition: analyzing induced EEG gamma-band responses in brain's source space. PloS One *2*, e684.

Supp, G.G., Siegel, M., Hipp, J.F., and Engel, A.K. (2011). Cortical hypersynchrony predicts breakdown of sensory processing during loss of consciousness. Curr. Biol. *21*, 1988–1993.

Sutton, S., Braren, M., Zubin, J., and John, E.R. (1965). Evoked-potential correlates of stimulus uncertainty. Science *150*, 1187–1188.

Synek, V.M. (1988). Prognostically important EEG coma patterns in diffuse anoxic and traumatic encephalopathies in adults. J. Clin. Neurophysiol. *5*, 161–174.

Tagliazucchi, E., and Laufs, H. (2014). Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. Neuron *82*, 695–708.

Tagliazucchi, E., von Wegner, F., Morzelewski, A., Brodbeck, V., Jahnke, K., and Laufs, H. (2013). Breakdown of long-range temporal dependence in default mode and attention networks during deep sleep. Proc. Natl. Acad. Sci. U. S. A. *110*, 15419–15424.

Tervaniemi, M., Maury, S., and Näätänen, R. (1994). Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity. Neuroreport *5*, 844–846.

Thut, G., Schyns, P.G., and Gross, J. (2011). Entrainment of Perceptually Relevant Brain Oscillations by Non-Invasive Rhythmic Stimulation of the Human Brain. Front. Psychol. *2*, 170.

Timofeev, I., Grenier, F., Bazhenov, M., Sejnowski, T.J., and Steriade, M. (2000). Origin of slow cortical oscillations in deafferented cortical slabs. Cereb. Cortex *10*, 1185–1199.

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

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

Tononi, G. (2012). Integrated information theory of consciousness: an updated account. Arch. Ital. Biol. *150*, 293–329.

Tononi, G. (2015). Integrated information theory. Scholarpedia *10*, 4164.

Tononi, G., and Edelman, G.M. (1998). Consciousness and complexity. Science *282*, 1846–1851.

Tononi, G., and Koch, C. (2008). The neural correlates of consciousness: an update. Ann. N. Y. Acad. Sci. *1124*, 239–261.

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? Philos. Trans. R. Soc. Lond. B. Biol. Sci. *370*, 20140167.

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

Tononi, G., Sporns, O., and Edelman, G.M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc. Natl. Acad. Sci. U. S. A. *91*, 5033–5037.

Tononi, G., Sporns, O., and Edelman, G.M. (1996). A complexity measure for selective matching of signals by the brain. Proc. Natl. Acad. Sci. U. S. A. *93*, 3422–3427.

Tononi, G., McIntosh, A.R., Russell, D.P., and Edelman, G.M. (1998a). Functional clustering: identifying strongly interactive brain regions in neuroimaging data. NeuroImage *7*, 133–149.

Tononi, G., Srinivasan, R., Russell, D.P., and Edelman, G.M. (1998b). Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. Proc. Natl. Acad. Sci. U. S. A. *95*, 3198–3203.

Tononi, G., Edelman, G.M., and Sporns, O. (1998c). Complexity and coherency: integrating information in the brain. Trends Cogn. Sci. *2*, 474–484.

Tzovara, A., Simonin, A., Oddo, M., Rossetti, A.O., and De Lucia, M. (2015). Neural detection of complex sound sequences in the absence of consciousness. Brain *138*, 1160–1166.

Uva, L., Librizzi, L., Wendling, F., and de Curtis, M. (2005). Propagation dynamics of epileptiform activity acutely induced by bicuculline in the hippocampal-parahippocampal region of the isolated Guinea pig brain. Epilepsia *46*, 1914–1925.

Vialatte, F.-B., Maurice, M., Dauwels, J., and Cichocki, A. (2010). Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. Prog. Neurobiol. *90*, 418–438.

Virtanen, J., Ruohonen, J., Näätänen, R., and Ilmoniemi, R.J. (1999). Instrumentation for the measurement of electric brain responses to transcranial magnetic stimulation. Med. Biol. Eng. Comput. *37*, 322–326.

Vitanyi, P.M.B., and Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. IEEE Trans. Inf. Theory *46*, 446–464.

Vuilleumier, P., Assal, F., Blanke, O., and Jallon, P. (2000). Distinct behavioral and EEG topographic correlates of loss of consciousness in absences. Epilepsia *41*, 687–693.

Wallace, C.S., and Dowe, D.L. (1999). Minimum Message Length and Kolmogorov Complexity. Comput. J. *42*, 270–283.

Walter, W.G., Cooper, R., Aldridge, V.J., McCALLUM, W.C., and Winter, A.L. (1964). Contingent negative variation : an electric sign of sensori-motor association and expectancy in the human brain. Nature *203*, 380–384.

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. Nature *393*, 440–442.

Weiss, B., Clemens, Z., Bódizs, R., Vágó, Z., and Halász, P. (2009). Spatio-temporal analysis of monofractal and multifractal properties of the human sleep EEG. J. Neurosci. Methods *185*, 116–124.

Wendling, F., Bellanger, J.J., Bartolomei, F., and Chauvel, P. (2000). Relevance of nonlinear lumped-parameter models in the analysis of depth-EEG epileptic signals. Biol. Cybern. *83*, 367–378.

Wendling, F., Bartolomei, F., Bellanger, J.J., and Chauvel, P. (2001). Interpretation of interdependencies in epileptic signals using a macroscopic physiological model of the EEG. Clin. Neurophysiol. *112*, 1201–1218.

Wendling, F., Ansari-Asl, K., Bartolomei, F., and Senhadji, L. (2009). From EEG signals to brain connectivity: a model-based evaluation of interdependence measures. J. Neurosci. Methods *183*, 9–18.

Wendt, H., Abry, P., and Jaffard, S. (2007). Bootstrap for empirical multifractal analysis. IEEE Signal Process. Mag. *24*, 38–48.

Wijnen, V.J.M., van Boxtel, G.J.M., Eilander, H.J., and de Gelder, B. (2007). Mismatch negativity predicts recovery from the vegetative state. Clin. Neurophysiol. *118*, 597–605.

Womelsdorf, T., Valiante, T.A., Sahin, N.T., Miller, K.J., and Tiesinga, P. (2014). Dynamic circuit motifs underlying rhythmic gain control, gating and integration. Nat. Neurosci. *17*, 1031–1039.

Wyart, V., and Tallon-Baudry, C. (2008). Neural dissociation between visual awareness and spatial attention. J. Neurosci. *28*, 2667–2679.

Yamamoto, Y., and Hughson, R.L. (1991). Coarse-graining spectral analysis: new method for studying heart rate variability. J. Appl. Physiol. *71*, 1143–1150.

Yu, F., Jiang, Q., Sun, X., and Zhang, R. (2015). A new case of complete primary cerebellar agenesis: clinical and imaging findings in a living patient. Brain *138*, e353.

Zanin, M., Zunino, L., Rosso, O.A., and Papo, D. (2012). Permutation Entropy and its main biomedical and econophysics applications: a review. Entropy *14*, 1553–1577.

Zellner, A. (1996). An Introduction to Bayesian Inference in Econometrics (New York; Chichester: Wiley-Interscience).

Zenil, H., Soler-Toscano, F., Dingle, K., and Louis, A.A. (2014). Correlation of automorphism group size and topological properties with program-size complexity evaluations of graphs and complex networks. Phys. Stat. Mech. Its Appl. *404*, 341–358.

Zenil, H., Kiani, N.A., and Tegnér, J. (2016). Methods of information theory and algorithmic complexity for network biology. Semin. Cell Dev. Biol. *51*, 32–43.

Zhang, Z. (1995). A fast method to compute surface potentials generated by dipoles within multilayer anisotropic spheres. Phys. Med. Biol. *40*, 335–349.

Zhang, X.S., Roy, R.J., and Jensen, E.W. (2001). EEG complexity as a measure of depth of anesthesia for patients. IEEE Trans. Biomed. Eng. *48*, 1424–1433.

Zhao, P., Van-Eetvelt, P., Goh, C., Hudson, N., Wimalaratna, S., and Ifeachor, E. (2007). Characterization of EEGs in Alzheimer's disease using information theoretic methods. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. *2007*, 5127–5131.

Zilber, N., Ciuciu, P., Abry, P., and Wassenhove, V. van (2012). Modulation of scale-free properties of brain activity in MEG. In 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1531–1534.

Zilber, N., Ciuciu, P., Abry, P., and Van Wassenhove, V. (2013). Learning-induced modulation of scale-free properties of brain activity measured with MEG. In 2013 IEEE 10th International Symposium on Biomedical Imaging, (IEEE), pp. 998–1001.

Ziv, J., and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. IEEE Trans. Inf. Theory *24*, 530–536.

7 Appendix A
KT: model building and compression

Model building and compression are best formalized through the concept of algorithmic or **Kolmogorov complexity** (the length of the shortest program capable of generating a string or dataset), further discussed below.

In order to clarify that the concepts of classification, compression, modeling and prediction are actually closely related if not equivalent, we note the following:

- If an agent has access to a good model for I/Os, then it can use it to compress it.
- If an agent has access to a good model for the I/Os, it can use it for memory and prediction of future data and planning around it.
- Classification is in fact an implicit form of modeling and can be used for data compression as well.

In what follows we equate consciousness with conscious awareness (Tononi and Koch, 2008) and, e.g., do not require self-awareness. We have argued that what we call reality is represented by the simplest programs our brains can find to model interaction with the environment. With IIT we now link the experience of reality, the qualia and phenomenal structure we associate with it to the phenomenology of consciousness (phenomenal states and structure). We have argued here that what we call reality is represented and shaped by the simplest programs our brains can find to model our interaction with the environment. In some sense, simplicity is equivalent to reality, and therefore experience. We propose the following hypothesis, relating cognition and consciousness:

Hypothesis 1. *Conscious states are experienced by agents running successful, simple models of their input/output streams. The more compressive these models are, the stronger the subjective experiences generated.*

An implicit element here is that consciousness is a graded and multidimensional phenomenon, it comes in shades and colors. In a sense, consciousness and awareness of reality will be associated to information processing systems that are efficient in describing and interacting with the external information world to some extent, and which can contrast prediction and reality (via feedback). An ant, for example, represents such a system. In fact, the distinction between life, cognition and consciousness appears more as a matter of degree.

Next, we discuss some consequences of this hypothesis. In (Ruffini, 2009), we hypothesized the following related conjecture with regard to the experience of Presence, which we may now view a consequence of Hypothesis 1 if we further assume that model selection (behavior) is driven by what feels most real:

Consequence 1. *Given alternate models (interpretations) for a given input/output information stream, an agent will select the simplest one it can construct that describes the data and agrees prior models (i.e., compressed prior data).*

That is, the feeling of Presence (a strong form of qualia, which we extend here more generally to awareness or the experience of particular qualia), as measured by subjective

or objective ways, is increased if the input/output data stream has an inherent low complexity, i.e., it can be modeled in a simple manner by the subjects brain compressing subsystems. Physical consistency in the inputs, in the sense of there being a simplifying low level model available to match the data, is an important element to enhance this experience, as it connects with low level, small memory capacity modeling mechanisms. As we progress higher in the modeling hierarchy, Bayesian prior expectations play an important role: explanations with a better match with past experiences (past data) are inherently simpler.

Another consequence of our hypothesis is the following, applicable in the ERP oddball paradigm (further discussed below) to attempt to quantify consciousness level:

Consequence 2. *Conscious experience (conscious level) is stronger (higher) in agents capable of identifying more complex patterns in their I/Os streams.*

In particular, a more conscious brain should be able to generate stronger responses to integrated coherent (compressible) input/output streams (e.g., sound and light from a common, coherent source, or the combination of input/output).

Given a spatio-temporal pattern of electric fields in the brain, how are the ones produced by successful compressive systems different than others? How can we distinguish brain chatter from that associated to some chaotic dynamical system? This is a key question in our program. We can draw inspiration from the work done with systems such as elementary cellular automata as discussed above. We provide methods to approximate K below. We use the notation of K_e for such estimates.

Consequence 3. *The level of consciousness can be mathematically characterized from signals generated by brains. Conscious brains create apparently compressible but ultimately compressible data streams.*

Here we note that a) proxies for complexity may be available but will be limited in practice, b) apparent complexity is not a monotonic function of conscious state. Since brain state is ultimately (Kolmogorov) compressible, both a very high or very low apparently complexity brain state should correlate with low consciousness.

For example, let us consider the PCI metric. Within the framework of K, it is reasonable to assume that LZW uncompressibility of spatially extended measurements of activity could be associated to integrated apparent complexity. If a system encodes a tight (Kolmogorov simple) model (e.g., let us picture here a model encoded in an RNN), we expect a strong response from perturbation of a node. The perturbations should have rather diverse effects at other nodes, but not in a totally random way since they are tied together by a model. Again, we would predict that there is an optimal regime of apparent complexity, too much or too little reflecting low consciousness.

Finally, we consider the concepts of conditional and **mutual algorithmic information/complexity**:

Consequence 4. *Consider a set of input data streams and the response of an agent as measured by neuroimaging data or agent behavior. A more conscious agent processing these inputs will lead to a lower algorithmic complexity of brain state or behavior data conditioned on the input data, or equivalently, the mutual algorithmic information of input data and response will be higher than in a less conscious state.*

The hypothesis is that the **mutual algorithmic information will be higher if subjects are conscious and attending.** If demonstrated, it could be used in clinical settings in which self-reports are not available (e.g., LIS or fetal consciousness). However, in more

precise terms, note that high mutual algorithmic information will also be present with the state of the optical nerve or cerebellum if we record data there during the duration of the external input. Our hypothesis is that information will be compressed in the cortex, represented as a model in the brain. Models are implemented in synaptic connectivity together with neuronal dynamics. Compressed spatiotemporal representations of the input streams should thus be present in neuroimaging data. It is in this sense that we expect mutual information between world and observer to increase with consciousness.

8 Appendix B
Description of four classical approaches used to estimate distributed dipole sources

Minimum Norm Estimate (MNE)

Minimum norm estimates (Dale and Sereno 1993; Hämäläinen and Ilmoniemi 1994) are based on a search for the solution with minimum power using the L2 norm to regularize the problem. This type of estimators is well appropriate to distributed source models where the dipole activity is likely to extend over some areas of the cortical surface.

$$\hat{D}_{MNE} = (G^T G + \lambda I)^{-1} G^T S$$

where I is the identity matrix and λ is the regularization parameter that weights the influence of priors in the solution.

Weighted Minimum Norm Estimate (wMNE)

The weighted Minimum Norm Estimate algorithm compensates for the tendency of MNE to favor weak and surface sources. This is done by introducing a weighting matrix W_S:

$$\hat{D}_{wMNE} = (G^T W_S G + \lambda I)^{-1} G^T W_S S$$

where matrix W_S adjusts the properties of the solution by reducing the bias inherent to MNE solutions. Classically, Ws is a diagonal matrix built from matrix G with non-zero terms inversely proportional to the norm of the lead field vectors.

Low resolution Brain Electromagnetic Tomography (LORETA)

In Low resolution electromagnetic tomography (LORETA) the main feature is to hypnotize that neighbor dipoles are strongly correlated. Therefore, a spatial smoothness constraint is explicitly promoted by applying a Laplacian operator to the sources in the regularization term. Moreover, as in wMNE, the columns of **G** are normalized to compensate for the misestimating of deep sources. In this method, for constrained number, position and orientation of dipolar sources (normal to the cortical surface), the estimate of the dipole moments is given by:

$$\hat{D}_{LORETA} = (G^T G + \lambda B \Delta^T \Delta B)^{-1} G^T S$$

where B is a diagonal matrix for the column normalization of **G** and Δ is a Laplacian operator.

Standardized low resolution brain electromagnetic tomography (sLORETA)

Despite its name, sLORETA (Pascual-Marqui, 2002) is not based on LORETA but rather on MNE. Indeed, sLORETA uses the source distribution estimated from MNE and standardizes it a posteriori by the variance of each estimated dipole source

$$\hat{D}_{sLORETA} = \hat{D}^{T}_{MNE,l} \left\{ \left| V_{\hat{D}} \right|_{ll} \right\}^{-1} \hat{D}_{MNE,l}$$

where $\hat{D}^{T}_{MNE,l}$ is the current density estimate at the *l*th voxel given by the minimum norm estimate and $\left\{ \left| V_{\hat{D}} \right|_{ll} \right\}$ is the *l*th diagonal block of the resolution matrix $V_{\hat{D}}$ (variance of the estimated current density) defined as $G^{T}\left[GG^{T} + \lambda I\right]^{-1}$. Therefore, contrarily to LORETA, MNE wMNE, sLORETA does not estimate intensity of a given source, but rather the probability of this source to disclose high amplitude as compared to the other ones.

LORETA and sLORETA inverse methods have been originally described using the whole brain volume as source space (Pascual-Marqui, 2002). For the present study, in order to ease the comparison with MNE and wMNE, we have implemented these methods by restricting the source space to the cortical surface.

The choice of λ is important and many approaches have been proposed to estimate it. Although there is no agreement on any optimal solution (David et al., 2002), as the main purpose of our study is to compare different approaches based on both inverse solutions and connectivity estimates, we chose to limit the number of intrinsic factors. Consequently, we used the same value of λ =1 for the four inverse algorithms based on the signal to noise ratio of our signals.

9 Appendix C
Steps of cortical sources estimation and methodological issues in EEG source connectivity

First step: estimation of cortical sources

Several approaches have been proposed to solve the inverse problem and these have been widely used in the context of brain source localization either in normal or pathological conditions. Among the methods especially designed for distributed brain sources, the most popular algorithms include (but are not limited to) the Minimum Norm Estimate (MNE) and its weighted version (wMNE) (Eddy et al., 2006; Hämäläinen and Ilmoniemi, 1994; Lin et al., 2004; Ruffini, 2015), Low resolution brain electromagnetic tomography (LORETA) and standardized low resolution brain electromagnetic tomography (sLORETA) (Pascual-Marqui, 2002; Pascual-Marqui et al., 1994). In addition, some efforts have been done to evaluate inverse algorithms in the view of localizing the brain sources in specific applications (Bai et al., 2007; Grova et al., 2006).

The EEG inverse problem can be stated as follows. According to the linear discrete equivalent current dipole model, EEG signals S(t) measured from M channels can be expressed as linear combinations of P time-varying current dipole sources D(t):

$$\mathbf{S} = G.\,\mathbf{D} + \mathbf{N}$$

where **G** and **N**(t) are respectively the matrix containing the lead fields of the dipolar sources and the additive noise.

In the general case, the inverse problem consists in finding an estimate $\hat{\mathbf{D}}(t)$ of the dipolar source parameters (typically, the position, orientation and magnitude), given the EEG signals **S**(t) and given the gain matrix **G.** This matrix can be computed from a multiple layer head model (volume conductor) and from the position of electrodes. For instance, the Boundary Element Method is a numerical method classically used in the case of realistic head models, but other methods are possible (see., e.g., (Ruffini, 2015), where we used realistic head models from FEM, see *Figure 9-1*).

As this problem is ill-posed (P>>M), physical and mathematical constraints have to be added to obtain a unique solution among the many solutions that minimize the residual term in the fitting of measured EEG signals.

For instance, using segmented MRI data, the source distribution can be constrained to a field of current dipoles homogeneously distributed over the cortex (Dale and Sereno, 1993), and normal to the cortical surface.

Technically, in the source model, we assume that EEG signals are generated by macro-columns of pyramidal cells lying in the cortical mantle and aligned orthogonally with respect to its surface (Nunez and Srinivasan, 2006). Thus, the electrical contribution of each macro-column to scalp electrodes can be represented by a current dipole located at the center of gravity of each triangle of the 3D mesh and oriented normally to the triangle surface. Using this source space, the so-called distributed approaches (described

in *Appendix A*
KT: model building and compression) only estimate the moment of dipole sources.

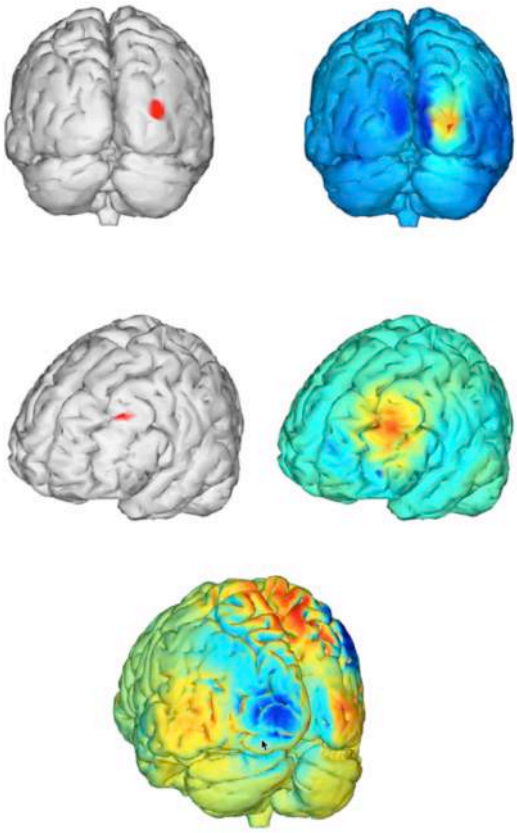

Figure 9-1. Sample cortical mapping using realistic head models (Ruffini, 2015). On the top and middle left, dipole sources, on the right reconstructed dipole fields from generated EEG data. On the bottom, cortical map of dipole field from a subjects EEG data in the alpha band (20 electrodes).

Second step: estimation of the functional connectivity

Regarding brain connectivity methods applied to EEG/MEG, bi- or multi-variate approaches proposed so far can be divided into two main categories depending on the assumptions made about the statistical coupling between signals. The first category includes linear methods such as the linear cross-correlation (R^2) (Brazier and Casby, 1952) or the coherence function (Brazier, 1968). The second category includes nonlinear methods based on mutual information (MI) (Mars and Lopes da Silva, 1983), nonlinear regression (h^2) (Pijn and Silva, 1993; Wendling et al., 2001), generalized synchronization (GS) (Stam et al., 2003) and phase synchronization (PS) (Rosenblum and Pikovsky, 2001). Recently, some efforts have been made to evaluate the performance of some of these measures in different scopes. An evaluation of the connectivity methods conducted on different synthetic and physiological models highlighted the high variability in the results provided by these methods, depending on the model that generates analyzed signals (Ansari-Asl et al., 2006; Wendling et al., 2009). Interestingly, a conclusion from these studies is that regression-based methods show the best performance, as opposed to more sophisticated methods which may be

blind to coupling changes. Finally, methods were also reported to reduce the effect of source leakage on functional connectivity measures. The general idea is to assume that a zero (or even very low) time lag is likely to correspond to a "volume conducted" activity as opposed to a "functionally related" activity for which a delay is expected due to synaptic transmission for instance. A method based on the imaginary part of the coherence function, often referred to as Imaginary Coherence (ImC), was initially proposed by Nolte et al.(Nolte et al., 2004) and was further improved regarding the bias inherent to the coherence estimation (Drakesmith et al., 2013). In section 4.2.1, we have already provided equations for nonlinear regression (h^2), mutual information *(MI)* and phase synchronization (*PLV)* which are also typical methods that be applied to the time-course of brain sources reconstructed from scalp EEG recordings.

Methodological issues in EEG source connectivity methods

The source connectivity approach has received increasing attention over the past decade (Astolfi et al., 2004, 2005; Babiloni et al., 2005; Brookes et al., 2011; David et al., 2002, 2002, 2003, 2003, 2004; Ghuman et al., 2011; Hoechstetter et al., 2004; Ray et al., 2007; Supp et al., 2007), see (Schoffelen and Gross, 2009) for review.

It is conceptually very appealing as networks are directly identified in the source space, typically in the neocortex. However, it raises a number of methodological issues. First, it requires i) to solve the ill-posed EEG/MEG inverse problem. Second, a functional connectivity method must be chosen among the many available ones. Third, although source connectivity methods tend to reduce volume conduction effects, these can never be completely cancelled in source space.

In a recent study (Hassan et al., 2014), we analyzed the impact of three factors that intervene in this processing: i) the number of scalp electrodes, ii) the combination between the algorithm used to solve the EEG inverse problem and the algorithm used to measure the functional connectivity and iii) the frequency bands retained to estimate the functional connectivity among neocortical sources. Using dense-EEG recordings in healthy volunteers, we evaluated these factors on evoked responses during a picture naming task. Results are illustrated in *Figure 9-2*. They show that the three studied factors have a dramatic impact on the final result (the identified network in the source space) as strong discrepancies were evidenced depending on the methods used. They also suggest that the combination of weighted Minimum Norm Estimator (wMNE) and the Phase Synchronization (PLV) methods applied on High-Resolution EEG in beta/gamma bands provides the best performance in term of topological distance between the EEG-identified network and the expected network, as reported in the literature, for the same task and based on fMRI.

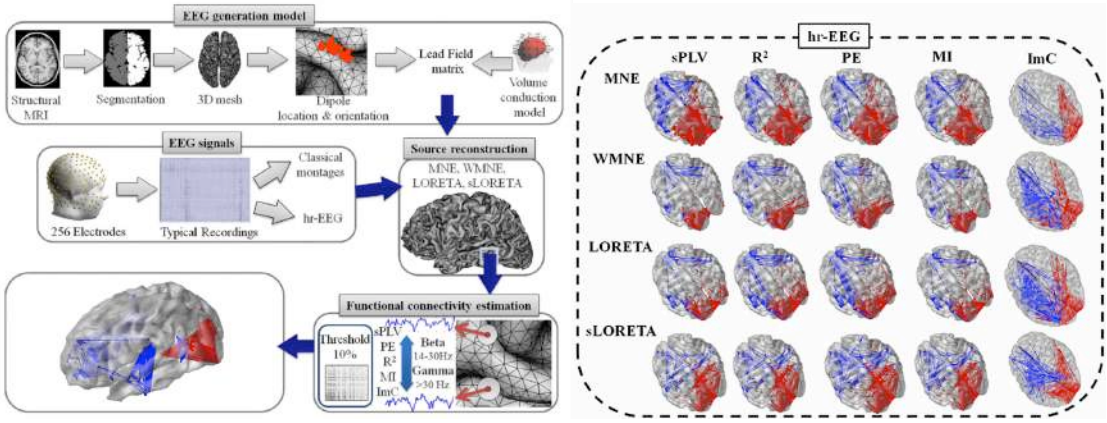

Figure 9-2. The different steps and the results of the comparative study. Left: a pipeline was developed to estimate neocortical brain networks from dense EEG data. Four inverse methods and five functional connectivity methods were combined. Right: results indicate that the topology of identified networks strongly depend on the chosen inverse method and, to a lesser extent, on the connectivity method. The combination wMNE+PLV was found to provide good results in term of distance between the EEG-identified network and fMRI-based results published on the same task (picture naming). Red and blue lines denote the functional connectivity as measured in the gamma (> 30 Hz) and beta (14-30 Hz) frequency band respectively. Abbreviations: hr-EEG: high-resolution EEG, MNE: Minimum norm estimate, wMNE: Weighted Minimum norm estimate, LORETA: Low resolution Brain Electromagnetic Tomography, sLORETA: Standardized Low resolution Brain Electromagnetic Tomography, sPLV: single-trial Phase Locking Value, PE: Phase Entropy, R2: linear correlation coefficient, MI: Mutual Information, ImC: Imaginary Coherence. (Adapted from (Hassan et al., 2014)).