Open access • Posted Content • DOI:10.1101/2020.11.25.398081

A Synergistic Workspace for Human Consciousness Revealed by Integrated Information Decomposition — Source link

Andrea I. Luppi, Pedro A. M. Mediano, Fernando Rosas, Judith Allanson ...+10 more authors

Institutions: University of Cambridge, Imperial College London, University of Western Ontario, Trinity College, Dublin

Published on: 26 Nov 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Information integration and Cognitive architecture

A Synergistic Workspace for Human Consciousness Revealed by Integrated Information Decomposition

Andrea I. Luppi[a,b]*, Pedro A.M. Mediano[c], Fernando E. Rosas[d,e,f], Judith Allanson[b,g], John D. Pickard[b], Robin L. Carhart-Harris[d], Guy B. Williams[b,h], Michael M Craig[a,b], Paola Finoia[b], Adrian M. Owen[i], Lorina Naci[j], David K. Menon[a,h], Daniel Bor[c], Emmanuel A. Stamatakis[a,b]

[a]Division of Anaesthesia, School of Clinical Medicine, University of Cambridge, United Kingdom

[b]Department of Clinical Neurosciences, University of Cambridge, Cambridge, United Kingdom

[c]Department of Psychology, University of Cambridge, Cambridge, United Kingdom

[d]Center for Psychedelic Research, Department of Brain Science, Imperial College London, London, United Kingdom

[e]Data Science Institute, Imperial College London, London, United Kingdom

[f]Centre for Complexity Science, Imperial College London, London, United Kingdom

[g]Department of Neurosciences, Cambridge University Hospitals NHS Foundation, Addenbrooke's Hospital, Cambridge, United Kingdom

[h]Wolfson Brain Imaging Centre, University of Cambridge, Cambridge, United Kingdom

[i]Department of Psychology and Department of Physiology and Pharmacology, The Brain and Mind Institute, University of Western Ontario, London, Ontario, Canada

[j]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

[g]Division of Neurosurgery, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom

*Corresponding author; email address: al857@cam.ac.uk

Abstract

A central goal of neuroscience is to understand how the brain synthesises information from multiple inputs to give rise to a unified conscious experience. This process is widely believed to require integration of information. Here, we combine information theory and network science to address two fundamental questions: how is the human information-processing architecture functionally organised? And how does this organisation support human consciousness? To address these questions, we leverage the mathematical framework of Integrated Information Decomposition to delineate a cognitive architecture wherein specialised modules interact with a "synergistic global workspace," comprising functionally distinct gateways and broadcasters. Gateway regions gather information from the specialised modules for processing in the synergistic workspace, whose contents are then further integrated to later be made widely available by broadcasters. Through data-driven analysis of resting-state functional MRI, we reveal that gateway regions correspond to the brain's well-known default mode network, whereas broadcasters of information coincide with the executive control network. Demonstrating that this synergistic workspace supports human consciousness, we further apply Integrated Information Decomposition to BOLD signals to compute integrated information across the brain. By comparing changes due to propofol anaesthesia and severe brain injury, we demonstrate that most changes in integrated information happen within the synergistic workspace. Furthermore, it was found that loss of consciousness corresponds to reduced integrated information between gateway, but not broadcaster, regions of the synergistic workspace. Thus, loss of consciousness may coincide with breakdown of information integration by this synergistic workspace of the human brain. Together, these findings demonstrate that refining our understanding of information-processing in the human brain through Integrated Information Decomposition can provide powerful insights into the human neurocognitive architecture, and its role in supporting consciousness.

Introduction

Humans and other vertebrates rely on a centralised nervous system to process information from the environment, obtained from a wide array of sensory sources. However, information from different sensory sources must eventually be combined and integrated with the organism's memories and goals, in order to guide adaptive behaviour effectively (Varela et al., 2001). Indeed, theoretical and empirical work in cognitive neuroscience indicates that information processed in parallel by domain-specific sensory modules (Taylor et al., 2009) needs to be integrated within a multimodal "central executive" (Fodor, 1985)

Furthermore, influential theories in computational neuroscience have also proposed that global integration of information from diverse sources plays a fundamental role in relation to human consciousness. The influential Global Neuronal Workspace Theory (GNWT) focuses on the process by which specific neural information becomes available for conscious access, as occurring through the global integration induced by a "global workspace" (Baars, 2005; Dehaene and Changeux, 2011; Dehaene et al., 2011; Mashour et al., 2020). The contents of the global workspace are thought to be widely broadcast back to localised processors, thereby providing "experiential integration" of distributed cortical modules into a coherent whole (Mashour et al., 2020). Also highlighting the importance of integration, the prominent Integrated Information Theory (IIT) posits that the degree of consciousness in a system is determined by its "integrated information": the amount of intrinsic information generated by the dynamics of the system considered as a whole, over and above the information generated by the dynamics of its individual constituent parts (Tononi, 2004; Tononi et al., 2016). Thus, this notion of integrated information corresponds to the extent to which "the whole is greater than the sum of its parts" (Balduzzi and Tononi, 2008)

Therefore, leading theoretical accounts of consciousness converge on this point: consciousness critically depends on the capability for global integration across a network of differentiated modules (Cavanna et al., 2018). However, despite this partial agreement, it is important to note that the role GNWT and IIT assign to integration with respect to consciousness is functionally distinct. On the one hand, GNWT builds on the premise that conscious information is globally available for further cognitive processing, contrasting with unconscious information that is only available within local modules. Thus, information becomes consciously accessible only upon being broadcasted to the rest of the brain, and integration is viewed as a necessary - but not sufficient - prerequisite step on the way to

broadcasting. In contrast, IIT proposes a more fundamental identity between the system's integrated information and subjective experience, but without specifying a formal architecture for this process: that is, according to IIT any system that integrates information will thereby be conscious, regardless of its specific organisation. Seen under this light, it becomes apparent that IIT and GWT are actually addressing different aspects of consciousness, and their views of integration are different but potentially complementary.

A crucial step for carrying these approaches forward is to deepen our understanding of how the brain's information-processing architecture enables the integration of information. Crucially, our ability to make sense of any information-processing architecture is limited by our understanding of the information that is being processed. An elegant framework to account for information in distributed systems is provided by *Partial Information Decomposition* (PID) (Williams and Beer, 2010), which posits that not all information is the same: two sources can possess information about a given target that is either unique (each source provides independent information), redundant (the same information is provided by both sources) or synergistic (complementary information, available only when both sources are considered together). As an example, humans have two sources of visual information about the world i.e. two eyes. The information that is lost when one eye is closed is called the "unique information" of that source - information that cannot be obtained from the remaining eye. The information that one still has when one eye is closed is called "redundant information" - because it is information that is carried equally by both sources. This provides robustness and resilience: i.e. you can still see even after losing one eye. However, losing one eye also deprives you of stereoscopic information about depth. This information does not come from either eye alone: you need both, in order to perceive the third dimension. Therefore, this is called the "synergistic information" between the sources - the extra advantage that is derived from combining them.

By applying a generalisation of PID to timeseries data - known as Integrated Information Decomposition (ΦID) (Mediano et al., 2019a) - to functional and diffusion MRI data, our previous work identified a set of brain regions that constitute a "synergistic core" supporting higher-level cognitive functions in the human brain - which we believe may play the role of the brain's global workspace (Luppi et al., 2020a). This "synergistic global workspace" mainly comprises high-level prefrontal and parietal cortical regions whose network organisation is especially well-suited for global integration of information across the brain,

which contrast with the modular and redundant interactions observed in sensorimotor areas (Luppi et al., 2020a). Building on these findings, it is natural to ask whether this synergistic workspace is related to human consciousness. Furthermore, given that the views on information integration put forward by GNWT and IIT are potentially complementary, an important challenge to move the field forward is to leverage both accounts into a unified architecture that could explain empirical effects observed in neuroimaging data.

Therefore, in this paper we set out to address two fundamental questions of contemporary neuroscience:

1. How is the human information-processing workspace functionally organised? Specifically: What brain regions does it involve, and what are the roles of the two types of information integration proposed by GNWT and IIT within the workspace?
2. How are these information-processing roles related to human consciousness?

To address these questions, we study three resting-state fMRI datasets: (i) data from the Human Connectome Project (Van Essen et al., 2013); (ii) N=15 healthy volunteers who were scanned before and after general anaesthesia with the intravenous propofol as well as during post-anaesthetic recovery (Luppi et al., 2019); (iii) N=22 patients suffering from chronic disorders of consciousness (DOC) as a result of severe brain injury (Luppi et al., 2019).

Results

Identification of workspace gateways and broadcasters

By considering principles of distributed information-processing, one can introduce a functional taxonomy to differentiate the constituent elements of the global workspace according to their function. Specifically, we propose to divide the information-processing stream within a distributed system (such as the human brain) in three key stages: (i) gathering of information from multiple distinct modules into a workspace; (ii) integration of the gathered information within the workspace; and (iii) global information broadcasting to the rest of the brain. Furthermore, we propose that while all workspace regions are involved in stage (ii), they are differentially involved in stages (i) and (iii).

154 The existence of a synergistic workspace and these three processing stages can be seen as
155 emerging from a trade-off between performance and robustness that is inherent to distributed
156 systems. Theoretical work in cognitive science (Baars, 2005; Fodor, 1985) and the field of
157 distributed signal processing (Tsitsiklis, 1989; Veeravalli and Varshney, 2012) has long
158 recognised the computational benefits of combining multiple distinct processing streams.
159 However, having a single source of inputs to and outputs from the workspace introduces what
160 is known as a "single point of failure," which can lead to catastrophic failure in case of
161 damage or malfunction (Lever et al., 2013). Therefore, a natural solution is to have not a
162 single but multiple units dedicated to gathering and broadcasting information, respectively,
163 thereby forming a workspace that can be in charge of synthesising the results of peripheral
164 processing (Rosas et al., 2017).

165 Focusing on Stage (ii), our previous work (Luppi et al., 2020a) identified which regions of
166 the human brain are predominantly synergistic, and thus are most reliant on combining
167 information from other brain regions. The key signature of workspace regions is to have a
168 high prevalence of synergistic (compared to redundant) connections, and therefore the
169 synergy-rich regions reported in (Luppi et al., 2020a) are ideally poised as GNW candidates.
170 Here, we consider the architecture of the global workspace more broadly, and combine
171 Integrated Information Decomposition with graph-theoretical principles to bring insights
172 about processing stages (i) and (iii) (Figure 1). We term this proposal the "Synergy-Φ-
173 Redundancy" neurocognitive architecture (SAPHIRE).

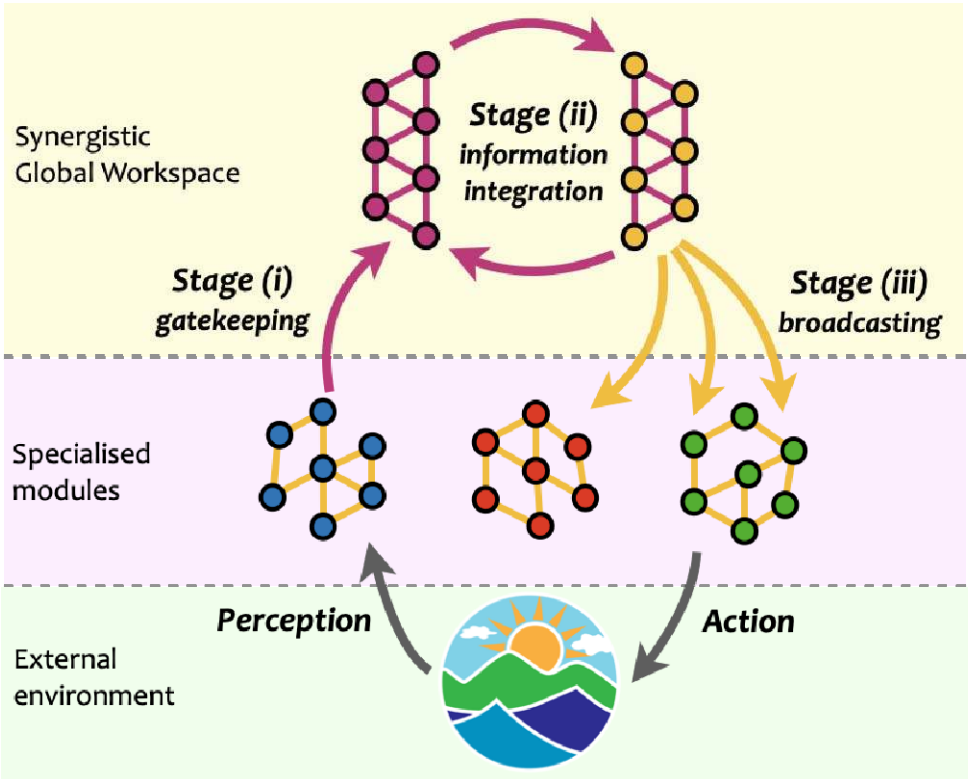

Figure 1. Schematic of the proposed SAPHIRE neurocognitive architecture. Below, specialised modules characterised by robust redundant interactions (Luppi et al., 2020a) process information about the environment. Information is then collected by workspace gateways through synergistic interactions [Step (i)]; synergistic interactions integrate information within the synergistic global workspace [Step (ii)]; workspace broadcasters spread the integrated information back to the specialised modules, through redundant interactions [Step (iii)], for further processing and to guide behaviour. Orange connections represent redundant interactions, and violet connections represent synergistic interactions. Grey connections represent interactions between the system and its environment.

We reasoned that brain regions through which information gains access to the workspace should exhibit synergistic connections that are widely distributed across the brain, as - by definition - the workspace gathers and synthesises information from a multiplicity of diverse brain modules. Thus, we postulate that regions that mediate the access to the synergistic workspace are connected with multiple modules within networks of synergistic interactions, synthesising incoming inputs from diverse sources (Sneve et al., 2019). We refer to such regions as *gateway*s (Figure 1, violet nodes). In contrast, the process of broadcasting information corresponds to disseminating multiple copies of the same information from the workspace to many functionally adjacent brain regions. Therefore, broadcaster regions also have connections with many different modules, but of non-synergistic, redundant

connections: "redundancy" accounts for the fact that multiple copies of the same information are being distributed. These regions are designated as *broadcasters* (Figure 1, orange nodes).

One approach to operationalise these ideas is by leveraging well-established graph-theoretical tools. In this work, we assess the diversity of intermodular connections using the *participation coefficient* (Rubinov and Sporns, 2010), which captures to what extent a given node connects to modules beyond its own (Methods). Note that this is different from the node *strength,* which captures a region's total amount of connectivity, and which we used to identify which regions belong to the synergistic workspace (Luppi et al., 2020a); the participation coefficient instead quantifies the diversity of modules that a region is connected to. Therefore, gateways are identified as brain regions that (a) belong to the workspace, and (b) have a highly-ranked participation coefficient in terms of synergistic connections. Conversely, broadcasters are global workspace regions that have a higher participation coefficient rank over redundant connections.

To explore these hypotheses, we computed synergistic and redundant interactions between 454 cortical and subcortical brain regions (Luppi and Stamatakis, 2020) based on resting-state functional MRI data from 100 subjects of the Human Connectome Project (following the same procedure as in Luppi *et al*, 2020 (Luppi et al., 2020a)). We then subdivided the brain into the well-established resting-state networks identified by Yeo and colleagues (Yeo et al., 2011), plus an additional subcortical module (Tian et al., 2020). Based on this partition into modules, we identified gateways and broadcasters by comparing the participation coefficients of synergistic versus redundant connections, for brain regions belonging to the synergistic workspace previously identified by Luppi *et al.,* (2020) (Luppi et al., 2020a) (Figure 2Ai,ii).

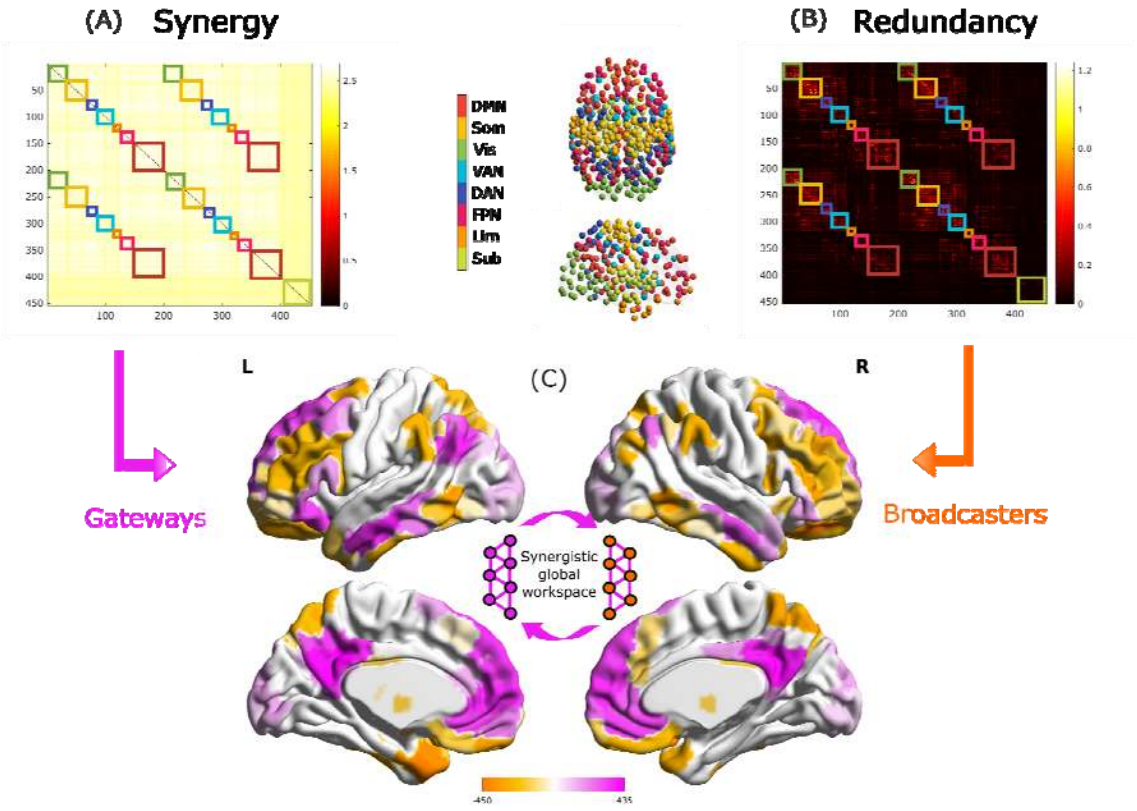

Figure 2. Gateways and broadcaster regions identified by their network connectivity profiles. (A) Group-average matrices of synergistic interactions between regions of the 454-ROI augmented Schaefer atlas(Luppi and Stamatakis, 2020; Schaefer et al., 2018). (B) Group-average matrices of redundant interactions. We highlighted modular allegiance to the canonical resting-state networks by using the colour scheme shown in between A and B. (C) Regions are identified as gateways (violet) or broadcasters (orange) based on the difference between rank of participation coefficient for synergy and redundancy, (only shown for brain regions identified as belonging to the synergistic global workspace by Luppi *et al* (Luppi et al., 2020a)). Violet indicates synergy rank > redundancy rank, corresponding to workspace regions that combine information of many brain modules (gateways); orange indicates the opposite, identifying workspace regions that broadcast information to many modules. Inset: illustration of the synergistic workspace. Legend: DMN, default mode network. Som, somatomotor network. Vis, visual network. VAN, ventral attention network. DAN, dorsal attention network. FPN, fronto-parietal control network. Lim, limbic network. Sub, subcortical network (comprised of 54 regions of the Tian 2020 atlas (Tian et al., 2020)). These results were also replicated using an alternative parcellation with 232 cortical and subcortical nodes (Supplementary Figure 1).

Intriguingly, our results reveal that gateways reside primarily in the brain's default mode network (Figure 2B, violet). In contrast, broadcasters are mainly located in the executive control network, especially lateral prefrontal cortex (Figure 2B, orange). Remarkably, the latter results are in line with Global Neuronal Workspace Theory, which consistently identifies lateral prefrontal cortex as a major broadcaster of information (Bor and Seth, 2012; Mashour et al., 2020).

Information decomposition identifies a synergistic core supporting human consciousness

Having introduced a taxonomy within the synergistic global workspace based on the distinct information-processing roles of different brain regions, we then sought to investigate their role in supporting human consciousness. Given the importance attributed to integration of information by both GNWT and IIT, we expected to observe reductions in integrated information within the areas of the synergistic workspace associated with loss of consciousness. Furthermore, we also reasoned that any brain regions that are specifically involved in supporting consciousness should "track" the presence of consciousness: the reductions should occur regardless of how loss of consciousness came about, and they should be restored when consciousness is regained.

We tested these hypotheses with resting-state fMRI from 15 healthy volunteers who were scanned before, during, and after anaesthesia with the intravenous agent propofol, as well as 22 patients with chronic disorders of consciousness (DOC) (Luppi et al., 2019). Resting-state fMRI data were parcellated into 400 cortical and 54 subcortical brain regions (Luppi and Stamatakis, 2020). Building on the IIT literature, which provides a formal definition of integrated information, we assessed integration corresponding to conscious activity via two alternative metrics: the well-known whole-minus-sum Φ measure introduced in (Balduzzi and Tononi, 2008), and the "revised Φ" (Φ-R) measure recently introduced in (Mediano et al., 2019a) (see Methods). Being demonstrably non-negative, this revised measure overcomes a major conceptual limitation of the original formulation of integrated information (Mediano et al., 2019a). For each subject, we computed the integrated information between each pair of BOLD signal timeseries, resulting in a 454-by-454 matrix of integrated information between brain regions. Treating this matrix as an (undirected) network enabled us to study consciousness-related changes in integrated information across conditions, which were analysed using the Network Based Statistic correction for multiple comparisons (Zalesky et al., 2010). Importantly, since we are interested in changes that are shared between the DOC and propofol datasets, we computed edge-level statistics using a composite null hypothesis test designed to detect such shared effects (see Methods).

Analysis based on Φ-R revealed a widespread reorganisation of integrated information throughout the brain when comparing awake volunteers against DOC patients, with both increases and decreases being observed ($p < 0.001$; Figure 3A). Likewise, propofol anaesthesia was also characterised by significant changes in integrated information between brain regions, both when compared with pre-anaesthetic wakefulness ($p < 0.001$; Figure 3B) and post-anaesthetic recovery ($p < 0.001$; Figure 3C).

Our analysis identified a number of the Φ-R connections that were reduced when consciousness was lost due to both anaesthesia and brain injury, and were restored during post-anaesthetic recovery - as we had hypothesised (Figure 3D). Remarkably, almost all regions showing consistent decreases in Φ-R when consciousness was lost were members of the global synergistic workspace, and specifically located in the default mode network (bilateral precuneus and medial prefrontal cortex) - and bilateral inferior parietal cortex – although left temporal cortices were also involved (Figure 3D). Additionally, some connections exhibited increases in Φ-R during loss of consciousness, and were restored upon recovery (Figure 3D), including areas in frontal cortex - especially lateral prefrontal cortex. Nevertheless, the overall balance was in favour of reduced integrated information: sum of F-scores associated with significant edges = -25.37 (Supplementary Figure 2). This was in contrast with the analysis based on the original formulation of Φ introduced by (Balduzzi and Tononi, 2008), which did not identify any reductions in integrated information that were common across anaesthesia and disorders of consciousness, instead only identifying common increases (Supplementary Figure 3A).

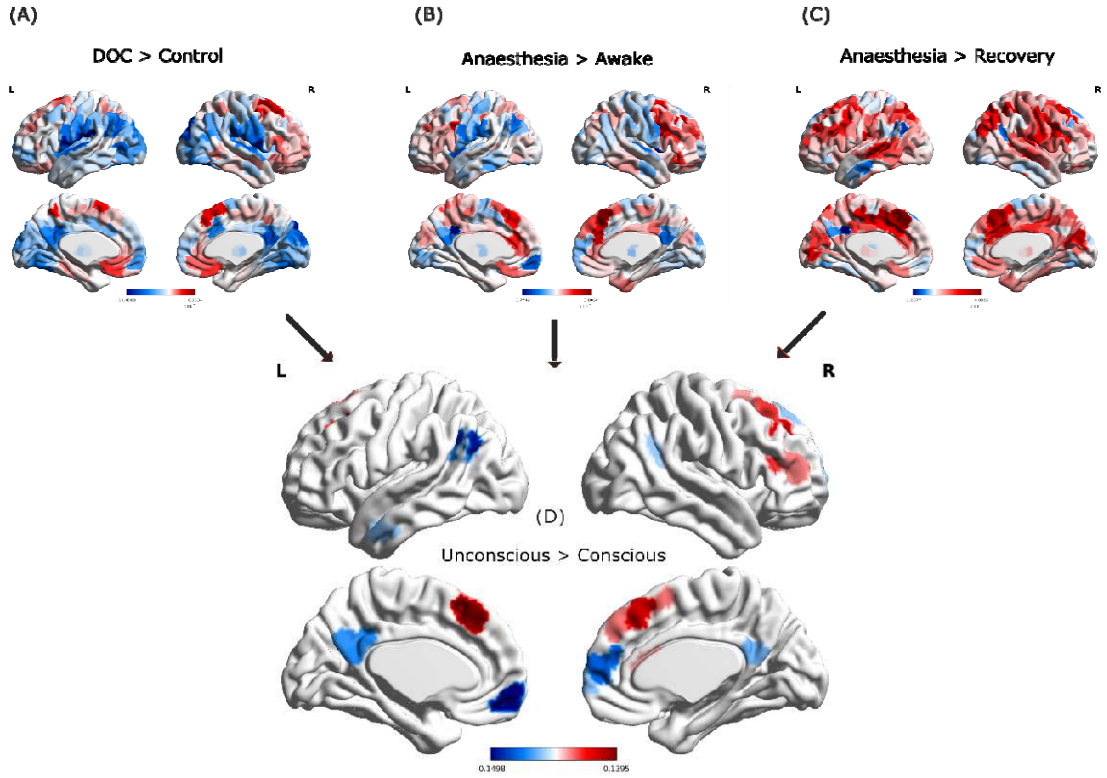

Figure 3. Loss of consciousness induces similar reorganisation of cortical integrated information across anaesthesia and disorders of consciousness. Top: Brain regions exhibiting overall NBS-corrected increases (red) and decreases (blue) in integrated information exchange when consciousness is lost. (A) DOC patients minus awake healthy volunteers; (B), propofol anaesthesia minus pre-induction wakefulness; (C) propofol-anaesthesia minus post-anaesthetic recovery. (D) Overlaps between the three contrasts in (A-C), showing increases and decreases that are common across anaesthesia and disorders of consciousness.

Having identified the subset of brain regions that are reliably associated with supporting human consciousness in terms of their integrated information, the last step of our analysis was to leverage the architecture proposed above to understand their role in the information processing stream within the brain. Since IIT predicts that loss of consciousness corresponds to reductions in integrated information, we focused on regions exhibiting reliable reductions in Φ-R when consciousness is lost (whether due to anaesthesia or DOC), which were restored upon recovery (shown in blue in Figure 3D).

Remarkably, our whole-brain results show that Φ-R disconnections induced by loss of consciousness play the role of gateway nodes (Figure 4A, violet) rather than broadcaster nodes (Figure 4A, orange) according to our previous identification (see Figure 2B, violet regions). Indeed, all reductions occur specifically within the default mode network (Figure

4B). Thus, these results suggest that loss of consciousness across anaesthesia and disorders of consciousness would correspond to anterior-posterior disconnection - in terms of integrated information - between DMN nodes that act as gateways into the synergistic workspace.

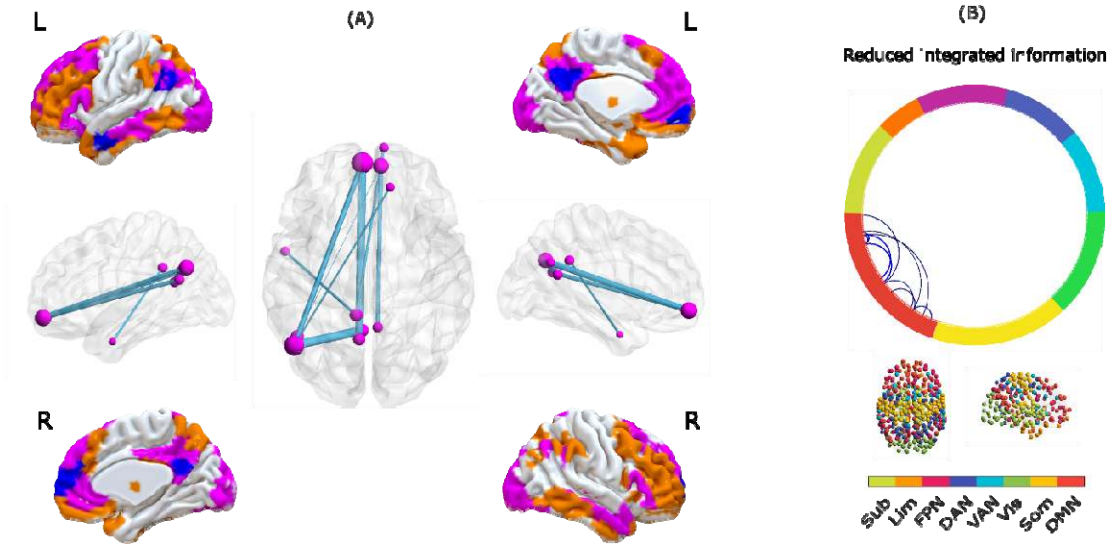

Figure 4. Synergistic core of human consciousness. (A) Surface projections indicate brain regions that play the role of broadcasters (orange) or gateways (violet) in the synergistic workspace, and regions that exhibit an overall significant reduction in integrated information across anaesthesia and disorders of consciousness (blue). Network representation: edges indicate reduced integrated information during both propofol anaesthesia and disorders of consciousness, between gateway (violet) and broadcaster (orange) nodes of the workspace. (B) Circular graph representation of significant reductions in integrated information (Φ-R) between brain regions, observed in all three contrasts, divided into canonical resting-state networks. Legend: DMN, default mode network. Som, somatomotor network. Vis, visual network. VAN, ventral attention network. DAN, dorsal attention network. FPN, fronto-parietal control network. Lim, limbic network. Sub, subcortical network (comprised of 54 regions of the Tian 2020 atlas (Tian et al., 2020)).

Replication with alternative parcellation

To ensure the robustness of our results to analytic choices, we also replicated them using an alternative cortical parcellation of lower dimensionality: we used the Schaefer scale-200 cortical parcellation (Schaefer et al., 2018), complemented with the scale-32 subcortical ROIs from the Tian subcortical atlas (Luppi and Stamatakis, 2020; Tian et al., 2020)

(Supplementary Figure 3B). Additionally, we also show that our results are not dependent on the choice of parameters in the NBS analysis, and are replicated using an alternative threshold definition for the connected component (extent rather than intensity) or a more stringent value for the cluster threshold ($F > 12$) (Supplementary Figure 3C-D). Importantly, whereas the increases in Φ-R are not the same across different analytic approaches, reductions of Φ-R in medial prefrontal and posterior cingulate/precuneus are reliably observed, attesting to their robustness.

Discussion

Architecture of the synergistic global workspace

This paper proposes a functional architecture for the brain's macroscale information processing flow, which leverages insights from network science and a refined understanding of neural information exchange. The synergy-Φ-redundancy (SAPHIRE) architecture posits the existence of a "synergistic workspace" of brain regions characterised by highly synergistic global interactions, which our previous work had shown to be composed by prefrontal and parietal cortices that are critical for higher cognitive functions (Luppi et al., 2020a). This workspace is further functionally decomposed by distinguishing gateways, which bring information from localised modules into the workspace, and broadcasters, which disseminate multiple copies of workspace information back to low-level regions.

Remarkably, our results on the HCP dataset show that the proposed operationalisation of gateways and broadcasters corresponds to the distinction between the brain's default mode network and executive control network, respectively. This data-driven identification of workspace gateways and broadcasters with the DMN and FPN provides a new framework to explain well-known functional differences between DMN and FPN, based on their distinct and complementary roles within the brain's synergistic global workspace, which is discussed below.

The fronto-parietal executive control network (FPN) mainly comprises lateral prefrontal and parietal cortices, and it is associated with performance of a variety of complex, cognitively demanding tasks (Barbey, 2018; Duncan and Owen, 2000; Fedorenko et al., 2013). A key

component of this network is lateral prefrontal cortex (LPFC). Based on theoretical and empirical evidence, as summarised in a recent review of GNWT (Mashour et al., 2020), this region is posited to play a major role in the global workspace, as a global broadcaster of information. Remarkably, this is precisely the role that our results assigned to LPFC, based on its combined information-theoretic and network properties. These results are also consistent with recent insights from network neuroscience, which indicate that the FPN is ideally poised to steer whole-brain dynamics through novel trajectories, in response to complex task demands (Barbey, 2018; Gu et al., 2015). Specifically, by broadcasting to the rest of the brain information that has been integrated within the workspace, the FPN may act as global coordinator of subsequent whole-brain dynamics.

On the other hand, the default mode network comprises posterior cingulate and precuneus, medial prefrontal cortex, and inferior parietal cortices (Fox et al., 2005; Raichle et al., 2001). Far from being merely a "task-negative network", as initially believed, the DMN is prominently involved in self-referential processing (Cavanna and Trimble, 2006; Qin and Northoff, 2011), and 'mental-time-travel' (Karapanagiotidis et al., 2017) or episodic memory and future-oriented cognition (Buckner and DiNicola, 2019; Buckner et al., 2008; Schacter et al., 2007; Szpunar et al., 2014). Its posterior regions in particular, act as relays between the neocortex and the hippocampal memory system (Buckner and DiNicola, 2019). This network was found to occupy a crucial position at the convergence of functional gradients of macroscale cortical organisation (Margulies et al., 2016), supporting its recently observed involvement in broader cognitive tasks (Vatansever et al., 2015b, 2015a, 2017). Thus, in terms of both neuroanatomical connectivity and functional engagement, the DMN is uniquely positioned to integrate and contextualise information coming into the global workspace (e.g. from sensory streams) by combining it with rich information pertaining to one's past experiences and high-level mental models about 'self' and world (Hassabis and Maguire, 2009) .

Integrated Information Decomposition of human consciousness

After identifying the neuroanatomical-functional mapping of the synergistic workspace in terms of gateways and broadcasters, we sought to identify their role in supporting human consciousness. Considering integrated information as a measure of consciousness, we

402 focused on identifying regions where information integration is reduced when consciousness
403 is lost (regardless of its cause, be it propofol anaesthesia or severe brain injury), and restored
404 upon its recovery. Our results indicate that brain regions exhibiting consciousness-specific
405 reductions in integrated information coincide with major nodes of the synergistic global
406 workspace.

407 Intriguingly, we found that the main disruptions of information integration were localised in
408 gateway nodes, rather than broadcasters. Thus, loss of consciousness in both anaesthesia and
409 disorders of consciousness could be understood as a breakdown of the entry points to the
410 "synergistic core" (Figure 4), which becomes unable to properly integrate inputs for the
411 workspace. Importantly, the original "whole-minus-sum" Φ introduced by Balduzzi and
412 Tononi (Balduzzi and Tononi, 2008) did not show consistent reductions during loss of
413 consciousness. Thus, the present results demonstrate the empirical validity of the "revised"
414 measure, Φ-R, in addition to its theoretical soundness (Mediano et al., 2019a).

415 Since workspace gateway regions coincide with the brain's default mode network, these
416 results are also in line with recent evidence that information content and integrative capacity
417 of the DMN are compromised during loss of consciousness induced by both anaesthesia and
418 severe brain injury (Boveroux et al., 2010; Hannawi et al., 2015; Luppi et al., 2019;
419 MacDonald et al., 2015; Di Perri et al., 2018; Vanhaudenhuyse et al., 2010). Due to its
420 prominent role in self-referential processing (Qin and Northoff, 2011), breakdown of DMN
421 connectivity within the synergistic workspace may be seen as a failure to integrate one's self-
422 narrative into the "stream of consciousness", in the words of Willam James.

423 This notion is further supported by focusing on reductions of integrated information during
424 anaesthesia compared with wakefulness. In addition to the synergistic core, overall reductions
425 are also observed in a set of thalamic, auditory and somatomotor regions, largely resembling
426 the brain regions that stop responding to sensory (auditory and noxious) stimuli once the
427 brain reaches propofol-induced saturation of EEG slow-wave activity (SWAS (Ní
428 Mhuircheartaigh et al., 2013)). Although there was no EEG data available to confirm this,
429 the doses of propofol employed in the present study are compatible with the doses of
430 propofol at which SWAS has been shown to arise (Warnaby et al., 2017), and therefore it is
431 plausible that our participants also reached SWAS and the loss of brain responsiveness it
432 indicates.

Thus, both resting-state integration of information between brain regions, as well as stimulus-evoked responses within each region (Ní Mhuircheartaigh et al., 2013), converge to indicate that propofol disrupts further processing of thalamocortical sensory information – a phenomenon termed "thalamocortical isolation" (Ní Mhuircheartaigh et al., 2013). We propose that as the thalamus and sensory cortices lose their ability to respond to stimuli, they cease to provide information to the synergistic core of the global workspace, resulting in a disconnection from the external world and presumably loss of consciousness.

These results testify to the power of the Integrated Information Decomposition framework: by identifying the information-theoretic components of integrated information, we have been able to obtain insights about human consciousness that remained elusive with alternative formulations, and could not be captured via standard functional connectivity or related methods. Thus, our findings support the notion that the global workspace is highly relevant for supporting consciousness in the human brain, in line with the proposal that *"[...] unconsciousness is not necessarily a complete suppression of information processing but rather a network dysfunction that could create inhospitable conditions for global information exchange and broadcasting"* (Mashour et al., 2020). GNWT postulates a key role for the global workspace in supporting consciousness: consistent with this theory, we find that several nodes of the synergistic global workspace become disconnected from each other in terms of integrated information when consciousness is lost, especially between anterior and posterior regions (Figure 4, brain networks). Thus, these are brain regions that (i) belong to the synergistic global workspace; (ii) exhibit overall reductions of integrated information when consciousness is lost; and (iii) are disconnected from other regions of the synergistic workspace when consciousness is lost. The brain regions satisfying these three conditions therefore constitute an interconnected "synergistic core" of workspace regions supporting human consciousness.

Limitations and future directions

Intriguingly, although we have focused on anaesthetic-induced decreases in integrated information, due to IIT's prediction that this is what should occur during loss of consciousness, our results also indicate concomitant increases of integrated information – possibly reflecting compensatory attempts (Figure 3). Interestingly, increases appear to

coincide with broadcaster nodes of the synergistic workspace. In particular, even though lateral prefrontal cortices are among the regions most closely associated with the global neuronal workspace in the literature (Bor and Seth, 2012; Mashour et al., 2020), our results indicate a paradoxical net increase in lateral prefrontal integrated information during anaesthesia and DOC. We interpret this qualitatively different behaviour as indicating that different subsets of the global workspace may be differentially involved in supporting consciousness. However, we note that, whereas the decreases in integrated information were robust to the use of different analytic approaches (e.g. use of a different parcellation or different NBS threshold), the increases that we observed were less robust, with no region consistently showing increases in integrated information (Supplementary Figure 3B-D). Nevertheless, both this phenomenon and the meaning of increased integrated information between brain regions deserve further investigation. Indeed, dreaming during anaesthesia has been reported to occur in up to 27% of cases (Leslie et al., 2007), and behaviourally unresponsive participants have been shown to perform mental imagery tasks during anaesthesia, both of which constitute cases of disconnected consciousness (Huang et al., 2018). Thus, although our doses of propofol were consistent with the presence of SWAS, we cannot exclude that some of our participants may have been merely disconnected but still conscious, possibly driving the increases we observed.

Therefore, in future work the use of independent measures to assess loss of consciousness, such as SWAS (Ní Mhuircheartaigh et al., 2013), the Perturbational Complexity Index (Casali et al., 2013), or quantification of the complexity of brain signals (Varley et al., 2020a) could provide stronger evidence that the brain changes we observed were actually due to unconsciousness rather than mere unresponsiveness. Additionally, it will be important to extend these results to other perturbations of consciousness: not only loss of consciousness induced by natural sleep or anaesthetics with different molecular mechanisms of action, such as the dissociative anaesthetic ketamine (Colombo et al., 2019; Li and Mashour, 2019; Mashour, 2014, 2016; Sarasso et al., 2015). More broadly, future research may also benefit from characterising the role of the synergistic workspace in the states of altered consciousness induced e.g. by psychedelics (Atasoy et al., 2018; Carhart-Harris, 2018; Luppi et al., 2020b), especially since prominent involvement of the DMN has already been identified (Carhart-Harris et al., 2014, 2016). Likewise, the use of paradigms different from resting-state, such as measuring the brain's spontaneous responses to engaging stimuli (e.g.

suspenseful narratives (Naci et al., 2017) or engaging movies (Naci et al., 2014)) may provide evidence for a more comprehensive understanding of brain changes during unconsciousness.

Additionally, the reliance here on 'resting-state' data without external stimuli may have resulted in an overestimation of the DMN's role in consciousness, and an under-estimation of the FPN (including lateral PFC), given their known different recruitment during no-task conditions (Fox et al., 2005). Indeed, recent efforts have been carried out to obtain a data-driven characterisation of the brain's global workspace based on regions' involvement across multiple different tasks (Deco et al., 2019). This work is complementary to ours in two aspects: first, the focus of (Deco et al., 2019) is on the role of the workspace related to cognition, whereas here we focus primarily on consciousness. Second, by using transfer entropy as a measure of functional connectivity, (Deco et al., 2019) assessed the directionality of information exchange – whereas our measure of integrated information is undirected, but are able to distinguish between different kinds of information being exchanged. Thus, different ways of defining and characterising a global workspace in the human brain are possible, and can provide complementary insights about distinct aspects of the human neurocognitive architecture.

It is also worth bearing in mind is that our measure of integrated information between pairs of regions does not amount to measuring the integrated information of the brain as a whole, as formally specified in the context of Integrated Information Theory (Balduzzi and Tononi, 2008; Tononi, 2004) - although we do show that the average integrated information between pairs of regions is overall reduced across the whole brain. We also note that our revised measure of integrated information is based on IIT 2.0 (Balduzzi and Tononi, 2008), which relies on a conceptually distinct understanding of integrated information from the most recent IIT 3.0 (Oizumi et al., 2014), whose computation requires perturbing the system and all of its subsets, making it computationally intractable. Thus, these limitations should be borne in mind when seeking to interpret the present results in the context of IIT.

We also acknowledge that our analyses did not include the cerebellum and brainstem, and future work may gain additional insights into their relevance for consciousness and cognition by extending our results to such regions. Finally, in order to obtain high spatial resolution, we relied here on brain signals based on the BOLD signal from functional MRI, which is only an indirect proxy of underlying neuronal activity, with limited temporal resolution. We sought to alleviate potential confounds of the hemodynamic response function by deconvolving it from

our data with a state-of-the-art toolbox (Wu et al., 2013) (Methods), which has been previously applied both in the context of anaesthesia (Wu et al., 2019), and of applying measures of integrated information to functional MRI data. Nevertheless, future applications of our analytic framework to M/EEG data may provide further insights.

Conclusion

Overall, we have shown that powerful insights about human neurocognitive architecture can be obtained through a more nuanced understanding of information exchange in the human brain, afforded by the framework of Integrated Information Decomposition.

Importantly, the proposed criteria to identify gateways, broadcasters, and the synergistic workspace itself, are based on practical network and information-theoretic tools, which are applicable to a broad range of neuroimaging datasets.

By refining and combining both the Global Neuronal Workspace Theory and Integrated Information Theory of consciousness, these findings bring us closer to a unified theoretical understanding of consciousness and its neuronal underpinnings - how mind arises from matter.

Materials and Methods

Anaesthesia Data

The propofol data employed in this study have been published before (Luppi et al., 2019; Naci et al., 2018; Varley et al., 2020a). For clarity and consistency of reporting, where applicable we use the same wording as our previous study (Luppi et al., 2019).

Recruitment

As previously reported (Luppi et al., 2019), "The propofol data were collected at the Robarts Research Institute in London, Ontario (Canada) between May and November 2014. A total of 19 (18–40 years; 13 males) healthy, right- handed, native English speakers, with no history of neurological disorders were recruited. Each volunteer provided written informed consent, following relevant ethical guidelines, and received monetary compensation for their time. The Health Sciences Research Ethics Board and Psychology Research Ethics Board of Western University (Ontario, Canada) ethically approved this study. Due to equipment malfunction or physiological impediments to anaesthesia in the scanner, data from three participants (1 male) were excluded from analyses, leaving 16" (Luppi et al., 2019).

Procedure

Resting-state fMRI data were acquired at no sedation (Awake), and Deep sedation (anaesthetised: Ramsay score of 5), and also during post-anaesthetic recovery. As previously reported (Luppi et al., 2019): "Ramsay level was independently assessed by two anaesthesiologists and one anaesthesia nurse in the scanning room before fMRI acquisition began, in each condition. Additionally, participants performed two tests: a computerised auditory target-detection task and a memory test of verbal recall, to evaluate their level of wakefulness independently of the assessors. For the Awake condition, participants did not receive a Ramsey score, as this scale is designed for patients in critical care. Instead, they had to be fully awake, alert and communicating appropriately. An infrared camera located inside the scanner was used to monitor wakefulness. For the Deep sedation condition, propofol was

administered intravenously using an AS50 auto syringe infusion pump (Baxter Healthcare, Singapore); step-wise sedation increments sedation were achieved using an effect-site/plasma steering algorithm combined with the computer-controlled infusion pump. Further manual adjustments were performed as required to reach target concentrations of propofol, as predicted by the TIVA Trainer (European Society for Intravenous Aneaesthesia, eurosiva.eu) pharmacokinetic simulation program. This software also specified the blood concentrations of propofol, following the Marsh 3-compartment model, which were used as targets for the pharmacokinetic model providing target-controlled infusion. The initial propofol target effect-site concentration was 0.6 μg mL^{-1}, with oxygen titrated to maintain SpO2 above 96%. Concentration was then increased by increments of 0.3 μg mL^{-1}, and Ramsay score was assessed: if lower than 5, a further increment occurred. Participants were deemed to have reached Ramsay level 5 once they stopped responding to verbal commands, were unable to engage in conversation, and were rousable only to physical stimulation. Data acquisition began once loss of behavioural responsiveness occurred for both tasks, and the three assessors agreed that Ramsay sedation level 5 had been reached. The mean estimated effect-site and plasma propofol concentrations were kept stable by the pharmacokinetic model delivered via the TIVA Trainer infusion pump; the mean estimated effect-site propofol concentration was 2.48 (1.82- 3.14) μg mL^{-1}, and the mean estimated plasma propofol concentration was 2.68 (1.92- 3.44) μg mL^{-1}. Mean total mass of propofol administered was 486.58 (373.30- 599.86) mg. These values of variability are typical for the pharmacokinetics and pharmacodynamics of propofol. At Ramsay 5 sedation level, participants remained capable of spontaneous cardiovascular function and ventilation. However, since the sedation procedure did not take place in a hospital setting, airway security could not be ensured by intubation during scanning, although two anaesthesiologists closely monitored each participant. Consequently, scanner time was minimised to ensure return to normal breathing following deep sedation. No state changes or movement were noted during the deep sedation scanning for any of the participants included in the study" (Luppi et al., 2019). Propofol was discontinued following the deep anaesthesia scan, and participants reached level 2 of the Ramsey scale approximately 11 minutes afterwards, as indicated by clear and rapid responses to verbal commands. This corresponds to the "recovery" period.

607

608 **Design**

609 As previously reported (Luppi et al., 2019): "In the scanner, subjects were instructed to relax
610 with closed eyes, without falling asleep; 8 minutes of fMRI scan without any task ("resting-
611 state") were acquired for each participant. Additionally, a separate 5-minute long scan was
612 also acquired while a plot-driven story was presented through headphones to participants,
613 who were instructed to listen while keeping their eyes closed" (Luppi et al., 2019). The
614 present analysis focuses on the resting-state data only; the story scan data have been
615 published separately (Kandeepan et al., 2020) and will not be discussed further here.

616 **Data Acquisition**

617 As previously reported (Luppi et al., 2019): "MRI scanning was performed using a 3-Tesla
618 Siemens Tim Trio scanner (32-channel coil), and 256 functional volumes (echo-planar
619 images, EPI) were collected from each participant, with the following parameters: slices = 33,
620 with 25% inter-slice gap; resolution = 3mm isotropic; TR = 2000ms; TE = 30ms; flip angle =
621 75 degrees; matrix size = 64x64. The order of acquisition was interleaved, bottom-up.
622 Anatomical scanning was also performed, acquiring a high-resolution T1- weighted volume
623 (32-channel coil, 1mm isotropic voxel size) with a 3D MPRAGE sequence, using the
624 following parameters: TA = 5min, TE = 4.25ms, 240x256 matrix size, 9 degrees FA" (Luppi
625 et al., 2019).

626

627 **Functional MRI preprocessing**

628 Following our previous work (Luppi et al., 2019), we preprocessed the functional imaging
629 data using a standard pipeline, implemented within the SPM12-based
630 (http://www.fil.ion.ucl.ac.uk/spm) toolbox CONN (http://www.nitrc.org/projects/conn),
631 version 17f(Whitfield-Gabrieli and Nieto-Castanon, 2012). As described, "The pipeline
632 comprised the following steps: removal of the first five scans, to allow magnetisation to reach
633 steady state; functional realignment and motion correction; slice-timing correction to account
634 for differences in time of acquisition between slices; identification of outlier scans for
635 subsequent regression by means of the quality assurance/artifact rejection software *Artifact*

636 *Detection Toolbox (art;* (http://www.nitrc.org/projects/artifact_detect); spatial normalisation
637 to Montreal Neurological Institute (MNI-152) standard space with 2mm isotropic resampling
638 resolution, using the segmented grey matter image from each volunteer's high-resolution T1-
639 weighted image, together with an *a priori* grey matter template" (Luppi et al., 2019).

640

641 **Disorders of Consciousness Patient Data**

642 The DOC patient functional data employed in this study have been published before (Luppi et
643 al., 2019, 2020b; Varley et al., 2020b). For clarity and consistency of reporting, where
644 applicable we use the same wording as our previous study (Luppi et al., 2019).

645

646 **Recruitment**

647 As previously reported (Luppi et al., 2019): "A sample of 71 DOC patients was included in
648 this study. Patients were recruited from specialised long-term care centres. To be invited to
649 the study, patients must have had a DOC diagnosis, written informed consent to participation
650 from their legal representative, and were capable of being transported to Addenbrooke's
651 Hospital. The exclusion criteria included any medical condition that made it unsafe for the
652 patient to participate (decision made by clinical personnel blinded to the specific aims of the
653 study) or any reason they are unsuitable to enter the MRI scanner environment (e.g. non-
654 MRI-safe implants), significant pre-existing mental health problems, or insufficient English
655 pre injury. After admission, each patient underwent clinical and neuroimaging testing.
656 Patients spent a total of five days (including arrival and departure days) at Addenbrooke's
657 Hospital. Coma Recovery Scale-Revised (CRS-R) assessments were recorded at least daily
658 for the five days of admission. If behaviours were indicative of awareness at any time,
659 patients were classified as MCS; otherwise UWS. We assigned MCS- or MCS+ sub-
660 classification if behaviours were consistent throughout the week. The most frequent signs of
661 consciousness in MCS- patients are visual fixation and pursuit, automatic motor reactions
662 (e.g. scratching, pulling the bed sheet) and localisation to noxious stimulation whereas MCS+
663 patients may, in addition, follow simple commands, intelligibly verbalise or intentionally but
664 inaccurately communicate [53,54]. Scanning occurred at the Wolfson Brain Imaging Centre,
665 Addenbrooke's Hospital, between January 2010 and December 2015; medication prescribed

to each patient was maintained during scanning. Ethical approval for testing patients was provided by the National Research Ethics Service (National Health Service, UK; LREC reference 99/391). All clinical investigations were conducted in accordance with the Declaration of Helsinki. As a focus of this study was on graph-theoretical properties of the brain, patients were systematically excluded from the final cohort analysed in this study based on the following criteria: 1) large focal brain damage (i.e. more than 1/3 of one hemisphere) as stated by an expert in neuroanatomy blinded to the patients' diagnoses; 2) excessive head motion during resting state scanning (i.e. greater than 3mm in translation and/or 3 degrees in rotation); 3) suboptimal segmentation and normalization of images. A total of 22 adults (14 males; 17 -70 years; mean time post injury: 13 months) meeting diagnostic criteria for Unresponsive Wakefulness Syndrome/Vegetative State or Minimally Conscious State due to brain injury were included in this study" (Luppi et al., 2019) (Table 1).

Table 1: Demographic information for patients with Disorders of Consciousness.

Sex	Age	Months post injury	Aetiology	Diagnosis	CRS-R Score
M	46	23	TBI	UWS	6
M	57	14	TBI	MCS-	12
M	46	4	TBI	MCS	10
M	35	34	Anoxic	UWS	8
M	17	17	Anoxic	UWS	8
F	31	9	Anoxic	MCS-	10
F	38	13	TBI	MCS	11
M	29	68	TBI	MCS	10
M	23	4	TBI	MCS	7
F	70	11	Cerebral bleed	MCS	9
F	30	6	Anoxic	MCS-	9
F	36	6	Anoxic	UWS	8
M	22	5	Anoxic	UWS	7
M	40	14	Anoxic	UWS	7

F	62	7	Anoxic	UWS	7
M	46	10	Anoxic	UWS	5
M	21	7	TBI	MCS	11
M	67	14	TBI	MCS-	11
F	55	6	Hypoxia	UWS	12
M	28	14	TBI	MCS	8
M	22	12	TBI	MCS	10
F	28	8	ADEM	UWS	6

Conscious State; TBI, Traumatic Brain Injury; fMRI-, negative responders to mental imagery task; fMRI+, positive responders to mental imagery task; SMA, supplementary motor area; PPA, parahippocampal place area; PMC, pre-motor cortex;

FMRI Data Acquisition

As previously reported (Luppi et al., 2019): "Resting-state fMRI was acquired for 10 minutes (300 volumes, TR=2000ms) using a Siemens Trio 3T scanner (Erlangen, Germany). Functional images (32 slices) were acquired using an echo planar sequence, with the following parameters: 3 x 3 x 3.75mm resolution, TR = 2000ms, TE = 30ms, 78 degrees FA. Anatomical scanning was also performed, acquiring high-resolution T1-weighted images with an MPRAGE sequence, using the following parameters: TR = 2300ms, TE = 2.47ms, 150 slices, resolution 1 x 1 x 1mm".

Functional MRI preprocessing

Due to the presence of deformations caused by brain injury, rather than relying on automated pipelines, patients' brains were individually preprocessed using SPM12, with visual inspections after each step. Additionally, to further reduce potential movement artifacts, data underwent despiking with a hyperbolic tangent squashing function. This is the same procedure followed in our previous studies (Luppi et al., 2019, 2020b).

Human Connectome Project data

We used the same functional MRI data from 100 unrelated subjects of the Human Connectome Project as in our previous work (Luppi et al., 2020a), with the same preprocessing and denoising procedures, which are also described below.

The dataset of functional and structural neuroimaging data used in this work came from the Human Connectome Project (HCP, http://www.humanconnectome.org/), Release Q3. Per HCP protocol, all subjects gave written informed consent to the HCP consortium. These data contained fMRI and diffusion weighted imaging (DWI) acquisitions from 100 unrelated subjects of the HCP 900 data release (Van Essen et al., 2013). All HCP scanning protocols were approved by the local Institutional Review Board at Washington University in St. Louis.

HCP: Functional data acquisition and denoising

As previously reported (Luppi and Stamatakis, 2020; Luppi et al., 2020a): "The following sequences were used: Structural MRI: 3D MPRAGE T1-weighted, TR= 2400 ms, TE = 2.14 ms, TI = 1000 ms, flip angle = 8°, FOV= 224 × 224, voxel size = 0.7 mm isotropic. Two sessions of 15 min resting-state fMRI: gradient-echo EPI, TR= 720 ms, TE= 33.1 ms, flip angle = 52°, FOV= 208 × 180, voxel size = 2 mm isotropic. Here, we only used functional data from the first scanning session, in LR direction".

We used the minimally preprocessed images made available by the HCP Consortium (Glasser et al., 2013), and subsequently followed the same aCompCor denoising pipeline described above, which is the same as in our previous work (Luppi and Stamatakis, 2020; Luppi et al., 2020a).

Brain Parcellation

Brains were parcellated into 454 cortical and subcortical regions of interest (ROIs). The 400 cortical ROIs were obtained from the scale-400 version of the recent Schaefer local-global functional parcellation (Schaefer et al., 2018). Since this parcellation only includes cortical regions, it was augmented with 54 subcortical ROIs from the highest resolution of the recent

Tian 2020 parcellation (Tian et al., 2020). We refer to this 454-ROI parcellation as the "augmented Schaefer" (Luppi and Stamatakis, 2020).

To ensure the robustness of our results to the choice of atlas, we also replicated them using an alternative cortical parcellation of different dimensionality: we used the Schaefer scale-200 cortical parcellation, complemented with the scale-32 subcortical ROIs from the Tian subcortical atlas (Luppi and Stamatakis, 2020).

BOLD timeseries extraction and HRF deconvolution

To construct matrices of functional connectivity, the timecourses of denoised BOLD signals were averaged between all voxels belonging to a given atlas-derived ROI, using the CONN toolbox. The resulting region-specific timecourses of each subject were then extracted for further analysis in MATLAB version 2016a.

IN accordance with our previous work (Luppi et al., 2020a) and previous studies using of information-theoretic measures in the context of functional MRI data, we used a state-of-the-art toolbox (Wu et al., 2013) to deconvolve the hemodynamic response function from our regional BOLD signal timeseries.

Integrated Information Decomposition

The framework of integrated information decomposition (ΦID) unifies integrated information theory (IIT) and partial information decomposition (PID) to decompose information flow into interpretable, disjoint parts. In this section we provide a brief description of ΦID and formulae required to compute the results in Figures 2 and 3. For further details, see (Mediano et al., 2019a).

Partial information decomposition

We begin with Shannon's Mutual information (MI), which quantifies the interdependence between two random variables X and Y. It is calculated as

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) \quad (1)$$

where $H(X)$ stands for the Shannon entropy of a variable X. Above, the first equality states that the mutual information is equal to the reduction in entropy (i.e. uncertainty) about X after Y is known. Put simply, the mutual information quantifies the information that one variable provides about another (Cover and Thomas, 2005).

Crucially, Williams and Beer (Williams and Beer, 2010) observed that the information that two source variables X and Y give about a third target variable Z, $I(X,Y ; Z)$, should be decomposable in terms of different *types* of information: information provided by one source but not the other (unique information), by both sources separately (redundant information), or jointly by their combination (synergistic information). Following this intuition, they developed the *Partial Information Decomposition* (PID; (Williams and Beer, 2010)) framework, which leads to the following fundamental decomposition:

$$I(X,Y;Z) = Red(X,Y;Z) + Un(X;Z|Y) + Un(Y;Z|X) + Syn(X,Y;Z). \quad (2)$$

Above, *Un* corresponds to the unique information one source but the other doesn't, *Red* is the redundancy between both sources, and *Syn* is their synergy: information that neither X nor Y alone can provide, but that can be obtained by considering X and Y together.

The simplest example of a purely synergistic system is one in which X and Y are independent fair coins, and Z is determined by the exclusive-OR function $Z = XOR(X,Y)$: i.e, Z=0 whenever X and Y have the same value, and Z=1 otherwise. It can be shown that X and Y are both statistically independent of Z, which implies that neither of them provide - by themselves - information about Z . However, X and Y together fully determine Z, hence the relationship between Z with X and Y is purely synergistic.

Recently, Mediano et al (2019) (Mediano et al., 2019a) formulated an extension of PID able to decompose the information that multiple source variables have about multiple target variables. This makes PID applicable to the dynamical systems setting, and yields a decomposition with redundant, unique, and synergistic components in the past and future that can be used as a principled method to analyse information flow in neural activity.

Synergy and redundancy calculation

As we previously observed (Luppi et al., 2020a): "While PID provides a formal framework, it does not enforce how the corresponding parts ought to be calculated. While there is ongoing research on the advantages of different decompositions for discrete data, most decompositions converge into the same simple form for the case of continuous Gaussian variables (Barrett, 2015). Known as *minimum mutual information PID* (MMI-PID), this decomposition quantifies redundancy in terms of the minimum mutual information of each individual source with the target; synergy, then, becomes identified with the additional information provided by the weaker source once the stronger source is known. Since linear-Gaussian models are sufficiently good descriptors of functional MRI timeseries (and more complex, non-linear models offer no advantage (Schulz et al., 2019)), here we adopt the MMI-PID decomposition, following previous applications of PID to neuroscientific data (Bím et al., 2019).

In a dynamical system such as the brain, one can calculate the amount of information flowing from the system's past to its future, known as time-delayed mutual information (TDMI). Specifically, by denoting the past of variables as $X_{t-\tau}$ and $Y_{t-\tau}$ and treating them as sources, and their joint future state (X_t, Y_t), as target, one can apply the PID framework and decompose the information flowing from past to future as

$$\begin{aligned} I(X_{t-\tau}, Y_{t-\tau};\, X_t, Y_t\,) \\ &= Red(X_{t-\tau}, Y_{t-\tau};\, X_t, Y_t\,) + Un(X_{t-\tau}; X_t, Y_t | Y_{t-\tau}) \\ &+ Un(Y_{t-\tau}; X_t, Y_t | X_{t-\tau}) + Syn(X_{t-\tau}, Y_{t-\tau};\, X_t, Y_t\,) \end{aligned} \tag{3}$$

Applying ΦID to this quantity allows us to distinguish between redundant, unique, and synergistic information shared with respect to the future variables X_t, Y_t (Mediano et al., 2019a). Importantly, this framework, has identified $Syn(X_{t-\tau}, Y_{t-\tau};\, X_t, Y_t\,)$ with the capacity of the system to exhibit emergent behaviour (Rosas et al., 2020) as well as a stronger notion of redundancy, in which information is shared by X and Y in both past and future. Accordingly, using the MMI-ΦID decomposition for Gaussian variables, we use

$$\mathrm{Red}(X, Y) = \min\{\, I(X_{t-\tau};\, X_t), I(X_{t-\tau};\, Y_t), I(Y_{t-\tau};\, X_t), I(Y_{t-\tau};\, Y_t)\}$$

$$\mathrm{Syn}(X,Y) = I(X_{t-\tau}, Y_{t-\tau};\ X_t, Y_t) - max\{ I(X_{t-\tau};\ X_t, Y_t), I(Y_{t-\tau};\ X_t, Y_t)\}. \quad (4)$$

$$\mathrm{Syn}(X,Y) = I(X_{t-\tau}, Y_{t-\tau};\ X_t, Y_t) - max\{ I(X_{t-\tau};\ X_t, Y_t), I(Y_{t-\tau};\ X_t, Y_t)\}. \quad (5)$$

Here, we used the Gaussian solver implemented in the JIDT toolbox (Lizier, 2014) to obtain TDMI, synergy and redundancy between each pair of brain regions, based on their HRF-deconvolved BOLD signal timeseries" (Luppi et al., 2020a).

Revised measure of integrated information

Through the framework of Integrated Information Decomposition (ΦID) we can decompose the constituent elements of Φ, the formal measure of integrated information proposed by Integrated Information Theory to quantify consciousness (Balduzzi and Tononi, 2008; Oizumi et al., 2014; Tononi, 2004). Note that several variants of Φ have been proposed over the years, including the original formulation of Tononi (Balduzzi and Tononi, 2008), other formulations based on causal perturbation (Oizumi et al., 2014) and others (see (Mediano et al., 2019b; Tegmark, 2016) for recent reviews). Here, we focus on the "empirical Φ" measure of (Barrett and Seth, 2011), based on the measures by Balduzzi and Tononi (2008) (Balduzzi and Tononi, 2008) and adapted to applications to experimental data. It is computed as

$$\Phi = I(X_{t-\tau}, Y_{t-\tau};\ X_t, Y_t) - I(X_{t-\tau};\ X_t) - I(Y_{t-\tau};\ Y_t) \quad (6)$$

and it quantifies how much temporal information is contained in the system over and above the information in its past. However, as others have pointed out (Oizumi et al., 2016), the measure in Eq. (6) can be negative, which is contradictory when taken as an absolute measure of integration.

Interestingly, with ΦID it can be shown that Φ is composed of different information atoms: it contains all the synergistic information in the system, the unique information transferred from X to Y and vice versa, and, importantly, a negative redundancy contribution - which explains why can be negative in redundancy-dominated systems.

To address this, Mediano et al. (2019) (Mediano et al., 2019a) introduced a revised measure of integrated information, Φ-R, which consists of the original Φ with the redundancy added back in,

$$\Phi\text{-R} = \Phi + \text{Red}(X, Y) \quad (7)$$

where Red(X, Y) is defined in Eq. (4). This measure is computationally tractable and preserves the original intuition of integrated information as measuring the extent to which "the whole is greater than the sum of its parts", since it captures only synergistic and transferred information. Crucially, thanks to Integrated Information Decomposition, it can be proved that the improved formulation of integrated information that we adopt here is guaranteed to be non-negative (Mediano et al., 2019a) - thereby avoiding a major conceptual limitation of the original formulation of Φ.

It is worth noting that more recent renditions of the theory exist (Oizumi et al., 2014). We do not consider the measure of integrated information proposed in IIT 3.0 (Oizumi et al., 2014) because it is computationally intractable for systems bigger than a small set of logic gates, and it is not universally well-defined (Barrett and Mediano, 2019).

Gradient of redundancy-to-synergy relative importance to identify the synergistic workspace

After building networks of synergistic and redundant interactions between each pair of regions of interest (ROIs), we determined the role of each ROI in terms of its relative engagement in synergistic or redundant interactions. Following the procedure previously described by Luppi et al (2020) (Luppi et al., 2020a), we first calculated the nodal strength of each brain region as the sum of all its connections in the group-averaged matrix. Then, we ranked all 454 regions based on their nodal strength (with higher-strength regions having higher ranks). This procedure was done separately for networks of synergy and redundancy. Subtracting each region's redundancy rank from its synergy rank yielded a gradient from negative (i.e. ranking higher in terms of redundancy than synergy) to positive (i.e. having a synergy rank higher than the corresponding redundancy rank); note that the sign is arbitrary.

It is important to note that the gradient is based on relative - rather than absolute - differences between regional synergy and redundancy; consequently, a positive rank difference does not necessarily mean that the region's synergy is greater than its redundancy; rather, it indicates that the balance between its synergy and redundancy relative to the rest of the brain is in favour of synergy - and *vice versa* for a negative gradient (Luppi et al., 2020a).

Subdivision of workspace nodes into gateways and broadcasters

To identify which regions within the workspace play the role of gateways or broadcasters postulated in our proposed architecture, we followed a procedure analogous to the one adopted to identify the gradient of redundancy-synergy relative importance (Luppi et al., 2020a), but replacing the node *strength* with the node *participation coefficient*.

The participation coefficient P_i quantifies the degree of connection that a node entertains with nodes belonging to other modules: the more of a node's connections are towards other modules, the higher its participation coefficient will be[8]. Conversely, the participation coefficient of a node will be zero if its connections are all with nodes belonging to its own module.

$$P_i = 1 - \sum_{s=1}^{M} \left(\frac{\kappa_{is}}{k_i} \right)^2 \quad (8)$$

Here, κ_{is} is the strength of positive connections between node i and other nodes in module s, k_i is the strength of all its positive connections, and M is the number of modules in the network. The participation coefficient ranges between zero (no connections with other modules) and one (equal connections to all other modules) [8].

Here, modules were set to be the seven canonical resting-state networks identified by Yeo and colleagues (Yeo et al., 2011), into which the Schaefer parcellation is already divided (Schaefer et al., 2018), with the addition of an eighth subcortical network comprising all ROIs of the Tian subcortical network (Tian et al., 2020). The brain's RSNs were chosen as modules because of their distinct and well-established functional roles, which fit well with the notion of modules as segregated and specialised processing systems interfacing with the

global workspace. Additionally, having the same definition of modules for synergy and redundancy allowed us to compute their respective participation coefficients in an unbiased way.

Separately for connectivity matrices of synergy and redundancy, the participation coefficient of each brain region was calculated. Then, regions belonging to the synergistic workspace were ranked, so that higher ranks indicated higher participation coefficient. Finally, the redundancy-based participation coefficient rank of each workspace region was subtracted from its corresponding synergy-based participation coefficient rank.

This procedure yielded a gradient over workspace regions, from negative (i.e. having a more highly ranked participation coefficient based on redundancy than synergy) to positive (i.e. having a more highly ranked participation coefficient based on synergy than redundancy). Note that as before, the sign of this gradient is arbitrary, and it is based on relative rather than absolute difference. Workspace regions with a positive gradient value were classified as "gateways", since they have synergistic connections with many brain modules. In contrast, workspace regions with a negative value of the gradient - i.e. those whose redundancy rank is higher than their synergy rank, in terms of participation coefficient - were labelled as workspace "broadcasters", since they possess information that is duplicated across multiple modules in the brain.

Statistical Analysis

Network Based Statistic

The network-based statistic approach (Zalesky et al., 2010) was used to investigate the statistical significance of propofol-induced or DOC-induced alterations on the networks of synergistic and redundant connections. This nonparametric statistical method is designed to control the family-wise error due to multiple comparisons, for application to graph data. Connected components of the graph are identified from edges that survive an a-priori statistical threshold (F-contrast; here we set the threshold to an F-value of 9). In turn, the statistical significance of such connected components is estimated by comparing their topology against a null distribution of the size of connected components obtained from non-parametric permutation testing. This approach rejects the null hypothesis on a component-by-

925 component level, and therefore achieves superior power compared to mass-univariate
926 approaches (Zalesky et al., 2010).

927

928 **Testing for common effects across datasets**

929 Since we are interested in the neural basis of consciousness, it is crucial to find changes that
930 are common across datasets, to rule out possible propofol- or DOC-specific effects that are
931 not related to consciousness per se (Luppi et al., 2019). To this end, we employed a null
932 hypothesis significance test under the composite null hypothesis that *at least* one dataset
933 among those considered here has no effect. In other words, for the null hypothesis to be
934 rejected we demand that all comparisons exhibit non-zero effects.

935

936 As usual, the test proceeds by comparing an observed test statistic with a null distribution.
937 The test statistic is the minimum of the three F-scores obtained in the comparisons of interest
938 (DOC vs awake; anaesthesia vs awake; and anaesthesia vs recovery), and the null distribution
939 is sampled by randomly reshuffling exactly one dataset (picked at random) at a time and
940 recalculating the F-scores. By shuffling exactly one dataset (instead of all of them), we are
941 comparing the observed data against the "least altered" version of the data that is still
942 compatible with the null hypothesis. This is a type of *least favourable configuration* (LFC)
943 test (Lehmann and Romano, 2005), which is guaranteed to control the false positive rate
944 below a set threshold (here, 0.05). The details of this test will be described in a future
945 publication.

946 Common connectivity changes across the three states of consciousness were then identified
947 as edges that were either (i) increased in DOC compared with control; (ii) increased during
948 anaesthesia compared with wakefulness; and (iii) increased during anaesthesia compared with
949 post-anaesthetic recovery; or (i) decreased in DOC compared with control; (ii) decreased
950 during anaesthesia compared with wakefulness; and (iii) decreased during anaesthesia
951 compared with post-anaesthetic recovery.

952

953 ## Acknowledgements

The authors would like to thank all the participants for their contribution to this study. This work was supported by grants from the UK Medical Research Council [U.1055.01.002.00001.01 to AMO and JDP]; The James S. McDonnell Foundation [to AMO and JDP]; and the Canada Excellence Research Chairs program (215063 to AMO); the National Institute for Health Research (NIHR, UK), Cambridge Biomedical Research Centre and NIHR Senior Investigator Awards [to DKM], the Stephen Erskine Fellowship (Queens' College, Cambridge, to EAS), the L'Oreal-Unesco for Women in Science Excellence Research Fellowship to LN; the British Oxygen Professorship of the Royal College of Anaesthetists [to DKM] and the Gates Cambridge Trust (to AIL). PAM and DB are funded by the Wellcome Trust (grant no. 210920/Z/18/Z). FR is funded by the Ad Astra Chandaria foundation. The research was also supported by the NIHR Brain Injury Healthcare Technology Co-operative based at Cambridge University Hospitals NHS Foundation Trust and University of Cambridge. AMO and DKM are Fellows of the CIFAR Brain, Mind, and Consciousness Programme. Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The image of the environment embedded in Figure 1 is available under CC-BY 3.0 license (https://creativecommons.org/licenses/by/3.0/) from WebStockReview.net.

Author Contributions

AIL: conceived the study; analysed data; wrote first draft of the manuscript. PAM: conceived the study; contributed to data analysis and interpretation of results; reviewed and edited the manuscript. FR: contributed to data analysis and interpretation of results; reviewed and edited the manuscript. M.M.C.: contributed to data analysis. DKM: reviewed the manuscript. RCH: reviewed the manuscript and contributed to interpretation of results. DB: conceived the study ; reviewed and edited the manuscript. EAS: conceived the study; reviewed and edited the manuscript. P.F., G.B.W., J.A., J.D.P., A.M.O., L.N., D.K.M. and E.A.S. were involved in designing the original studies for which the present data were collected. P.F., M.M.C., G.B.W., J.A., L.N. and E.A.S. all participated in data collection.

Competing Interests

The authors declare no competing interests.

Data and Code Availability

The CONN toolbox is freely available online (http://www.nitrc.org/projects/conn).
The Brain Connectivity Toolbox code used for graph-theoretical analyses is freely available online (https://sites.google.com/site/bctnet/).
The Java Information Dynamics Toolbox is freely available online:
(https://github.com/jlizier/jidt).
The HRF deconvolution toolbox is freely available online:
(https://www.nitrc.org/projects/rshrf).
The Network Based Statistic code is freely available online:
https://www.nitrc.org/projects/nbs/
The HCP fMRI data are available online (https://www.humanconnectome.org/study/hcp-young-adult/data-releases).
The propofol and DOC patient data that support the findings of this study are available from Dr.. Emmanuel Stamatakis, University of Cambridge (email: eas46@cam.ac.uk) upon reasonable request.

References

Atasoy, S., Vohryzek, J., Deco, G., Carhart-harris, R.L., and Kringelbach, M.L. (2018). Common neural signatures of psychedelics: Frequency-specific energy changes and repertoire expansion revealed using connectome-harmonic decomposition. Prog. Brain Res. *242*.

Baars, B.J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Prog. Brain Res. *150*.

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

1014 Barbey, A.K. (2018). Network Neuroscience Theory of Human Intelligence. Trends Cogn. Sci. *22*, 8–
1015 20.

1016 Barrett, A.B. (2015). Exploration of synergistic and redundant information sharing in static and
1017 dynamical Gaussian systems. Phys. Rev. E *91*, 52802.

1018 Barrett, A.B., and Mediano, P.A.M. (2019). The phi measure of integrated information is not well-
1019 defined for general physical systems. J. Conscious. Stud. *26*, 11–20.

1020 Barrett, A.B., and Seth, A.K. (2011). Practical Measures of Integrated Information for Time-Series
1021 Data. PLoS Comput Biol *7*, 1001052.

1022 Bím, J., De Feo, V., Chicharro, D., Bieler, M., Hanganu-Opatz, I., Brovelli, A., and Panzeri, S.
1023 (2019). A Non-negative Measure Of Feature-Related Information Transfer Between Neural Signals.
1024 BioRxiv *doi: https*.

1025 Bor, D., and Seth, A.K. (2012). Consciousness and the prefrontal parietal network: Insights from
1026 attention, working memory, and chunking. Front. Psychol. *3*.

1027 Boveroux, P., Vanhaudenhuyse, A., and Phillips, C. (2010). Breakdown of within- and between-
1028 network Resting State during Propofol-induced Loss of Consciousness. Anesthesiology *113*, 1038–
1029 1053.

1030 Buckner, R.L., and DiNicola, L.M. (2019). The brain's default network: updated anatomy, physiology
1031 and evolving insights. Nat. Rev. Neurosci. *20*, 593–608.

1032 Buckner, R.L., Andrews-Hanna, J.R., and Schacter, D.L. (2008). The brain's default network:
1033 Anatomy, function, and relevance to disease. Ann. N. Y. Acad. Sci. *1124*, 1–38.

1034 Carhart-Harris, R.L. (2018). The entropic brain - revisited. Neuropharmacology *142*, 167–178.

1035 Carhart-Harris, R.L., Leech, R., Hellyer, P.J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo,
1036 D.R., and Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging
1037 research with psychedelic drugs. Front. Hum. Neurosci. *8*, 20.

1038 Carhart-Harris, R.L., Muthukumaraswamy, S., Roseman, L., Kaelen, M., Droog, W., Murphy, K.,
1039 Tagliazucchi, E., Schenberg, E.E., Nest, T., Orban, C., et al. (2016). Neural correlates of the LSD
1040 experience revealed by multimodal neuroimaging. Proc. Natl. Acad. Sci. *113*, 201518377.

1041 Casali, A.G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K.R., Casarotto, S., Bruno,
1042 M.-A., Laureys, S., Tononi, G., et al. (2013). A Theoretically Based Index of Consciousness
1043 Independent of Sensory Processing and Behavior. In Science Translational Medicine, pp. 1–10.

1044 Cavanna, A.E., and Trimble, M.R. (2006). The precuneus: A review of its functional anatomy and
1045 behavioural correlates. Brain *129*, 564–583.

1046 Cavanna, F., Vilas, M.G., Palmucci, M., and Tagliazucchi, E. (2018). Dynamic functional
1047 connectivity and brain metastability during altered states of consciousness. Neuroimage *180*, 383–
1048 395.

1049 Colombo, M.A., Napolitani, M., Boly, M., Gosseries, O., Casarotto, S., Rosanova, M., Brichant, J.F.,
1050 Boveroux, P., Rex, S., Laureys, S., et al. (2019). The spectral exponent of the resting EEG indexes the
1051 presence of consciousness during unresponsiveness induced by propofol, xenon, and ketamine.
1052 Neuroimage *189*, 631–644.

1053 Cover, T.M., and Thomas, J.A. (2005). Elements of Information Theory (Wiley-Interscience).

1054 Deco, G., Vidaurre, D., and Kringelbach, M.L. (2019). Revisiting the global workspace: Orchestration
1055 of the functional hierarchical organisation of the human brain. BioRxiv.

1056 Dehaene, S., and Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious
1057 Processing. Neuron *70*, 200–227.

1058 Dehaene, S., Changeux, J.P., and Naccache, L. (2011). The global neuronal workspace model of
1059 conscious access: From neuronal architectures to clinical applications. Res. Perspect. Neurosci. *18*,
1060 55–84.

1061 Duncan, J., and Owen, A.M. (2000). Common regions of the human frontal lobe recruited by diverse
1062 cognitive demands. Trends Neurosci. *23*, 475–483.

1063 Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., and Ugurbil, K. (2013). The
1064 WU-Minn Human Connectome Project: An overview. Neuroimage *80*, 62–79.

1065 Fedorenko, E., Duncan, J., and Kanwisher, N. (2013). Broad domain generality in focal regions of
1066 frontal and parietal cortex. Proc. Natl. Acad. Sci. U. S. A. *110*, 16616–16621.

1067 Fodor, J.A. (1985). Précis of *The Modularity of Mind*. Behav. Brain Sci. *8*, 1–5.

1068 Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., and Raichle, M.E. (2005). The
1069 human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl.
1070 Acad. Sci. *102*, 9673–9678.

1071 Glasser, M.F., Sotiropoulos, S.N., Wilson, A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J.,
1072 Jbabdi, S., Webster, M., Polimeni, J.R., et al. (2013). The Minimal Preprocessing Pipelines for the
1073 Human Connectome Project. Neuroimage *80*, 105–124.

1074 Gu, S., Pasqualetti, F., Cieslak, M., Telesford, Q.K., Yu, A.B., Kahn, A.E., Medaglia, J.D., Vettel,
1075 J.M., Miller, M.B., Grafton, S.T., et al. (2015). Controllability of structural brain networks. Nat.
1076 Commun. *6*.

1077 Hannawi, Y., Lindquist, M.A., Caffo, B.S., Sair, H.I., and Stevens, R.D. (2015). Resting brain activity
1078 in disorders of consciousness: a systematic review and meta-analysis. Neurology *84*, 1272–1280.

1079 Hassabis, D., and Maguire, E.A. (2009). The construction system of the brain. Philos. Trans. R. Soc.
1080 B Biol. Sci. *364*, 1263–1271.

1081 Huang, Z., Vlisides, P.E., Tarnal, V.C., Janke, E.L., Keefe, K.M., Collins, M.M., McKinney, A.M.,
1082 Picton, P., Harris, R.E., Mashour, G.A., et al. (2018). Brain imaging reveals covert consciousness
1083 during behavioral unresponsiveness induced by propofol. Sci. Rep. *8*, 1–11.

1084 Kandeepan, S., Rudas, J., Gomez, F., Stojanoski, B., Valluri, S., Owen, A.M., Naci, L., Nichols, E.S.,
1085 and Soddu, A. (2020). Modeling an auditory stimulated brain under altered states of consciousness
1086 using the generalized ising model. Neuroimage *223*, 117367.

1087 Karapanagiotidis, T., Bernhardt, B.C., Jefferies, E., and Smallwood, J. (2017). Tracking thoughts:
1088 Exploring the neural architecture of mental time travel during mind-wandering. Neuroimage *147*,
1089 272–281.

1090 Lehmann, E.L., and Romano, J.P. (2005). Testing Statistical Hypotheses Third Edition With 6
1091 Illustrations.

Leslie, K., Skrzypek, H., Paech, M.J., Kurowski, I., and Whybrow, T. (2007). Dreaming During Anesthesia and Anesthetic Depth in Elective Surgery Patients. Anesthesiology *106*, 33–42.

Lever, K.E., Merabti, M., and Kifayat, K. (2013). Single Points of Failure Within Systems-of-Systems. 14th Annu. Post Grad. Symp. Converg. Telecommun. Netw. Broadcast. 183–188.

Li, D., and Mashour, G.A. (2019). Cortical dynamics during psychedelic and anesthetized states induced by ketamine. Neuroimage *196*, 32–40.

Lizier, J.T. (2014). JIDT: An Information-Theoretic Toolkit for Studying the Dynamics of Complex Systems. Front. Robot. AI *1*, 1–37.

Luppi, A.I., and Stamatakis, E.A. (2020). Combining network topology and information theory to construct representative brain networks. Netw. Neurosci. 1–46.

Luppi, A.I., Craig, M.M., Pappas, I., Finoia, P., Williams, G.B., Allanson, J., Pickard, J.D., Owen, A.M., Naci, L., Menon, D.K., et al. (2019). Consciousness-specific dynamic interactions of brain integration and functional diversity. Nat. Commun. *10*.

Luppi, A.I., Mediano, P.A., Rosas, F.E., Holland, N., Fryer, T.D., O'Brien, J.T., Rowe, J.B., Menon, D.K., Bor, D., and Stamatakis, E.A. (2020a). A synergistic core for human brain evolution and cognition. BioRxiv 2020.09.22.308981.

Luppi, A.I., Vohryzek, Jakub, Kringelbach, M.L., Mediano, P.A., Craig, M.M., Adapa, R., Carhart-Harris, R.L., Roseman, L., Pappas, I., Finoia, P., Williams, G.B., et al. (2020b). Connectome Harmonic Decomposition of Human Brain Dynamics Reveals a Landscape of Consciousness. BioRxiv.

MacDonald, A.A., Naci, L., MacDonald, P.A., and Owen, A.M. (2015). Anesthesia and neuroimaging: Investigating the neural correlates of unconsciousness. Trends Cogn. Sci. *19*, 100–107.

Margulies, D.S., Ghosh, S.S., Goulas, A., Falkiewicz, M., Huntenburg, J.M., Langs, G., Bezgin, G., Eickhoff, S.B., Castellanos, F.X., Petrides, M., et al. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. Proc. Natl. Acad. Sci. U. S. A. *113*, 12574–12579.

Mashour, G.A. (2014). Top-down mechanisms of anesthetic-induced unconsciousness. Front. Syst. Neurosci. *8*.

Mashour, G.A. (2016). Network-level Mechanisms of Ketamine Anesthesia. Anesthesiology 873–888.

Mashour, G.A., Roelfsema, P., Changeux, J.P., and Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron *105*, 776–798.

Mediano, P.A.M., Rosas, F., Carhart-Harris, R.L., Seth, A.K., and Barrett, A.B. (2019a). Beyond integrated information: A taxonomy of information dynamics phenomena. ArXiv.

Mediano, P.A.M., Seth, A.K., and Barrett, A.B. (2019b). Measuring integrated information: Comparison of candidate measures in theory and simulation. Entropy *21*.

Naci, L., Cusack, R., Anello, M., and Owen, A.M. (2014). A common neural code for similar conscious experiences in different individuals. Proc. Natl. Acad. Sci. U. S. A. *111*, 14277–14282.

Naci, L., Sinai, L., and Owen, A.M. (2017). Detecting and interpreting conscious experiences in

behaviorally non-responsive patients. Neuroimage *145*, 304–313.

Naci, L., Haugg, A., MacDonald, A., Anello, M., Houldin, E., Naqshbandi, S., Gonzalez-Lara, L.E., Arango, M., Harle, C., Cusack, R., et al. (2018). Functional diversity of brain networks supports consciousness and verbal intelligence. Sci. Rep. *8*.

Ní Mhuircheartaigh, R., Warnaby, C., Rogers, R., Jbabdi, S., and Tracey, I. (2013). Slow-wave activity saturation and thalamocortical isolation during propofol anesthesia in humans. Sci. Transl. Med. *5*, 208ra148.

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*.

Oizumi, M., Tsuchiya, N., and Amari, S.I. (2016). Unified framework for information integration based on information geometry. Proc. Natl. Acad. Sci. U. S. A. *113*, 14817–14822.

Di Perri, C., Amico, E., Heine, L., Annen, J., Martial, C., Larroque, S.K., Soddu, A., Marinazzo, D., and Laureys, S. (2018). Multifaceted brain networks reconfiguration in disorders of consciousness uncovered by co-activation patterns. Hum. Brain Mapp. *39*, 89–103.

Qin, P., and Northoff, G. (2011). How is our self related to midline regions and the default-mode network? Neuroimage *57*, 1221–1233.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., and Shulman, G.L. (2001). A default mode of brain function. Proc. Natl. Acad. Sci. U. S. A. *98*, 676–682.

Rosas, F., Hsiao, J.H., and Chen, K.C. (2017). A technological perspective on information cascades via social learning. IEEE Access *5*, 22605–22633.

Rosas, F.E., Mediano, P.A.M., Jensen, H.J., Seth, A.K., Barrett, A.B., Carhart-Harris, R.L., and Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. ArXiv.

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. Neuroimage *52*, 1059–1069.

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.

Schacter, D.L., Addis, D.R., and Buckner, R.L. (2007). Remembering the past to imagine the future: the prospective brain. Nat. Rev. Neurosci. *8*, 657–661.

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., and Yeo, B.T.T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb. Cortex *28*, 3095–3114.

Schulz, M.-A., Yeo, B.T.T., Vogelstein, J., Mourao-Miranada, J., Kather, J., Kording, K., Richards, B., and Bzdok, D. (2019). Deep learning for brains?: Different linear and nonlinear scaling in UK Biobank brain images vs. machine-learning datasets. BioRxiv *5*, 16.

Sneve, M.H., Grydeland, H., Rosa, M.G.P., Paus, T., Chaplin, T., Walhovd, K., and Fjell, A.M. (2019). High-expanding regions in primate cortical brain evolution support supramodal cognitive flexibility. Cereb. Cortex *29*, 3891–3901.

1170 Szpunar, K.K., Spreng, R.N., and Schacter, D.L. (2014). A taxonomy of prospection: Introducing an
1171 organizational framework for future-oriented cognition. Proc. Natl. Acad. Sci. U. S. A. *111*, 18414–
1172 18421.

1173 Taylor, K.I., Stamatakis, E.A., and Tyler, L.K. (2009). Crossmodal integration of object features:
1174 Voxel-based correlations in brain-damaged patients. Brain *132*, 671–683.

1175 Tegmark, M. (2016). Improved Measures of Integrated Information. PLoS Comput. Biol. *12*.

1176 Tian, Y., Margulies, D., Breakspear, M., and Zalesky, A. (2020). Topographic organization of the
1177 human subcortex unveiled with functional connectivity gradients. Nat. Neurosci. *23*, 1421–1432.

1178 Tononi, G. (2004). An information integration theory of consciousness An information integration
1179 theory of consciousness. BMC Neurosci. *5*, 42–64.

1180 Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: From
1181 consciousness to its physical substrate. Nat. Rev. Neurosci. *17*, 450–461.

1182 Tsitsiklis, J.N. (1989). Decentralized Detection. In Advances in Statistical Signal Processing, Vol 2:
1183 Signal Detection, p.

1184 Vanhaudenhuyse, A., Quentin Noirhomme, Ã., Luaba J-F Tshibanda, Ã., Bruno, M.-A., Boveroux, P.,
1185 Schnakers, C., Soddu, A., Perlbarg, V., Ledoux, D., Brichant, J.-F., et al. (2010). Default network
1186 connectivity reflects the level of consciousness in non-communicative brain- damaged patients. Brain
1187 *133*, 161–171.

1188 Varela, F., Lachaux, J.P., Rodriguez, E., and Martinerie, J. (2001). The brainweb: Phase
1189 synchronization and large-scale integration. Nat. Rev. Neurosci. *2*, 229–239.

1190 Varley, T.F., Luppi, A.I., Pappas, I., Naci, L., Adapa, R., Owen, A.M., Menon, D.K., and Stamatakis,
1191 E.A. (2020a). Consciousness & Brain Functional Complexity in Propofol Anaesthesia. Sci. Rep. *10*.

1192 Varley, T.F., Craig, M., Adapa, R., Finoia, P., Williams, G., Allanson, J., Pickard, J., Menon, D.K.,
1193 and Stamatakis, E.A. (2020b). Fractal dimension of cortical functional connectivity networks &
1194 severity of disorders of consciousness. PLoS One *15*.

1195 Vatansever, D., Menon, D.K., Manktelow, A.E., Sahakian, B.J., and Stamatakis, E.A. (2015a).
1196 Default mode network connectivity during task execution. Neuroimage *122*, 96–104.

1197 Vatansever, D., Menon, X.D.K., Manktelow, A.E., Sahakian, B.J., and Stamatakis, E.A. (2015b).
1198 Default Mode Dynamics for Global Functional Integration. J. Neurosci. *35*, 15254–15262.

1199 Vatansever, D., Menon, D.K., and Stamatakis, E.A. (2017). Default mode contributions to automated
1200 information processing. Proc. Natl. Acad. Sci. U. S. A. *114*, 12821–12826.

1201 Veeravalli, V. V, and Varshney, P.K. (2012). Distributed inference in wireless sensor networks.
1202 Trans. R. Soc. A *370*, 100–117.

1203 Warnaby, C.E., Sleigh, J.W., Hight, D., Jbabdi, S., and Tracey, I. (2017). Investigation of Slow-wave
1204 Activity Saturation during Surgical Anesthesia Reveals a Signature of Neural Inertia in Humans.
1205 Anesthesiology *127*, 645–657.

1206 Whitfield-Gabrieli, S., and Nieto-Castanon, A. (2012). Conn: A Functional Connectivity Toolbox for
1207 Correlated and Anticorrelated Brain Networks. Brain Connect. *2*, 125–141.

1208 Williams, P.L., and Beer, R.D. (2010). Nonnegative Decomposition of Multivariate Information.

ArXiv.

Wu, G.R., Liao, W., Stramaglia, S., Ding, J.R., Chen, H., and Marinazzo, D. (2013). A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data. Med. Image Anal. *17*, 365–374.

Wu, G.R., Di Perri, C., Charland-Verville, V., Martial, C., Carrière, M., Vanhaudenhuyse, A., Laureys, S., and Marinazzo, D. (2019). Modulation of the spontaneous hemodynamic response function across levels of consciousness. Neuroimage *200*, 450–459.

Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zollei, L., Polimeni, J.R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. *106*, 1125–1165.

Zalesky, A., Fornito, A., and Bullmore, E.T. (2010). Network-based statistic: Identifying differences in brain networks. Neuroimage *53*, 1197–1207.