frontiers in
COMPUTATIONAL NEUROSCIENCE

IMPACT 2.5 FACTOR

Physically orthodox theories-of-consciousness must predict biophysical mechanisms for the integration of spatially-distributed and temporally-extended neural codes

Nicholas Rosseinsky

Journal Name:	Frontiers in Computational Neuroscience
ISSN:	1662-5188
Article type:	Hypothesis & Theory Article
First received on:	10 Mar 2014
Frontiers website link:	www.frontiersin.org

Physically orthodox theories-of-consciousness must predict biophysical mechanisms for the integration of spatially-distributed and temporally-extended neural codes

Nicholas M. Rosseinsky[1,*]

[1]Department of Neuroscience, Center for Dialog in Science, Exeter, U.K.

*** Correspondence:** Nicholas M. Rosseinsky, Department of Neuroscience, Center for Dialog in Science, 14 Rosebarn Avenue, Exeter, EX4 6DY, U.K.
rosseinsky.nicholas.m@gmail.com

Keywords: neural codes, biophysics, consciousness, locality, causality, distributed codes, temporal codes, binding problem.

Word count: 11,999

Abstract

The characterization of certain spatiotemporal brain dynamics as "encodings" *e.g.* of the external environment is central to modern neuroscience. It is usually assumed that *any* encoding scheme employed in the brain can be utilized as the informational basis for human conscious experience, thus supporting a default view that no "new physics" is required for the explanation of consciousness. Here, in contrast, it is shown that the assumption of orthodox physics severely *constrains* encoding schemes for conscious experience, specifically because physical orthodoxy requires local causality, and proposes that no physical structures beyond the brain itself are involved in the generation of consciousness. Under these conditions, spatially-distributed codes cannot be the final encoding basis for consciousness, despite their ubiquitous appearance in the current neuroscience of brain structure and function, because required aggregation of information encoded at distinct spatial locations would be definitively non-local. Similarly, rate codes and other, more complex, temporally-extended schemes are non-local in time. Thus, the assumption of physical orthodoxy limits the encoding basis for components-of-consciousness to instantaneous states at single locations. Consequently, if physically-orthodox theories are to explain consciousness, biophysical mechanisms must exist to reduce well-established spatially- and temporally-extended codes to spatiotemporally point-like states, thus creating a strong and specific prediction for the biophysical architecture of neural information processing in the human brain. Future empirical exclusion of required biophysical mechanisms would establish that physically-novel mechanisms govern the generation of consciousness. Conversely, future empirical discovery of required mechanisms would be a strong pointer to the physical basis of human consciousness, if these mechanisms have no other functional role in human biology.

1. Introduction

Understanding consciousness is sometimes characterized as one of the greatest outstanding scientific challenges (Miller, 2005). The mainstream approach (Anderson, 1972; Baars, 1988; Edelman, 1989; Crick and Koch, 1990; Tononi and Edelman, 1998; Tegmark, 2000; Churchland, 2005; Baars and Edelman, 2012) is that certain brain dynamics are more or less synonymous with conscious experience, thus reducing the challenge to the empirical characterization (Dehaene, 2014) of those dynamical states that are both necessary and sufficient for particular conscious experiences. In this mainstream view, consciousness is either identical to particular brain dynamics (Place, 1956; Smart, 1959; Tegmark, 2000) or a conventionally-emergent property (Anderson, 1972; Van Gulick, 2001; Chalmers, 2008), so that consciousness can be explained in a physically orthodox setting, without any need for "new physics" (Edelman, 1989). Given the success of modern physical theory in explaining other natural phenomena within a single conceptual framework, the assumption that consciousness is a physically orthodox phenomenon in this sense is perhaps a natural hypothesis. Nevertheless, at present, this hypothesis has no empirical support, with the result that alternative, non-orthodox, proposals [*e.g.* (Penrose, 1989; Stapp, 1993; Chalmers, 1996; Beck and Eccles, 2003)] can arguably claim a scientific status equivalent to that of the mainstream view.

The present paper is the first in a series of three (Rosseinsky, 2014a, 2014b) whose goal is to develop objective empirical tests in ordinary neuroscience that can either falsify the "physically orthodox" hypothesis or provide new and significant information concerning the physical basis of consciousness (in orthodox settings). Unlike the vast majority of consciousness-related work in neuroscience (Tononi and Koch, 2008) that focuses on relatively macroscopic hypotheses *e.g.* concerning brain areas (Merker, 2007; Shulman *et al*., 2009) or global phase relationships (Klimesch *et al*., 2001; Melloni *et al*., 2007), the present series focuses on *neural codes* that act as the informational basis of consciousness. Requirements for codes are necessarily requirements for neural wiring and associated biophysical machinery, and therefore result directly in empirically-observable predictions for the class of physically orthodox theories-of-consciousness. (Because these empirical tests examine the biophysical structure of the human brain, they do not depend at all on detailed subjective report.)

This paper examines the requirement from physical orthodoxy that theories-of-consciousness must obey *locality in physical causation*, and the consequent implications for brain-dynamical encoding of consciousness. Although exceptions to locality within orthodox physical theory are sometimes argued to occur in certain quantum-theoretic contexts, what is meant here by a "physically orthodox theory-of-consciousness" (Baars, 1988; Edelman, 1989; Crick and Koch, 1990; Tononi and Edelman, 1998; Churchland, 2005; Baars and Edelman, 2012) is a member of the class of mainstream explanations that are all based solely on classical physics. Thus, physically-orthodox theories-of-consciousness must be local, in present nomenclature.

For an introduction to locality issues (**Figure 1**), consider as an illustrative example the encoding of color at a particular location in the visual field, by the vector formed from the firing rates of three different neurons (Solomon and Lennie, 2007). One typical mainstream proposal (Churchland, 1985) is that the firing-rate vector is *identical to* or *synonymous with* the associated conscious experience of color. Although a hypothesis of this kind can be advanced, it is *not* physically orthodox, because it fails to be local in physical causation: the color-experience depends on information at three different spatial locations. This simple observation perhaps evokes a jump directly to the central result, that physically-orthodox (and therefore locally causal) theories-of-consciousness can only be valid if each piece of fundamental information employed in the generation of conscious experience is encoded by the brain in a spatiotemporally point-like manner. Clearly, however, such a leap would be premature, because the example fails to exhaust all possibilities for physically orthodox theories-of-consciousness. For example, consider the proposal that one of the emergent properties of the brain is a spatial coordinate system in which the three vector-components *can* be considered co-located. [Although this particular idea might appear esoteric, a complete and explicit account of space as it relates to conscious experience (Dennett, 1991; Dennett and Kinsbourne, 1992) is required both generally in the neuroscience of consciousness and specifically to treat locality properties rigorously.] These considerations raise the primary question addressed in this paper: under what circumstances can generation of consciousness from spatially-distributed codes satisfy the locality requirements of physical orthodoxy?

A complete answer to this question must address the related contexts of *distributed* neural codes and *parallel computation* (Rumelhart and McClelland, 1988; Thorpe and Imbert, 1989), because both distributed codes and parallel computation employ physical dynamics at spatially-separated locations to encode related information (**Figure 2**). In the present paper, a *distributed code* utilizes neural dynamics in a number of spatially distinct neurons to encode a *single* piece of fundamental information (for example, the specific orientation of an edge, or the specific hue of color, at a specific external location). (Conventionally, the terminology "local code" is used to describe the complementary case of single-neuron encoding. Because the term "local" will be used in this paper solely as a descriptor of physical causality, instead the nomenclature "non-distributed" or "spatially-point-like" will be used to describe single-neuron codes.) In contrast to the distributed/non-distributed distinction that refers to basic encoding of a *single* fundamental piece of information, the *parallel computation* nomenclature will be used here to refer to multiple distinct streams that encode or process *multiple* pieces of related information (for example, edge-orientations at a number of locations, or edge-orientation, color, and motion at a single location). Thus, in present terminology, neural codes can in principle be non-distributed and non-parallel, distributed and parallel, non-distributed and parallel, or distributed and parallel, although the human sensory encoding and information processing that will form the focus of central developments is understood (Kaas, 1997; Decharms and Zador, 2000) to be implemented, at least in early stages, via massively parallel computations that typically employ distributed codes.

Data encoded at different spatial locations in the brain must be spatially integrated at some point in processing, if physical action based on that information is to be locally causal. (Otherwise, brain dynamics at multiple distinct locations would have a physical effect on some non-connected spatial location, requiring definitively *non*-local causality.) Problems involved with the integration of spatially-distributed encoding have been discussed to some extent under the heading of "the binding problem" (von der Malsburg, 1981; Treisman, 1999; Engel and Singer, 2001; Di Lollo, 2012; Feldman, 2013), although these literatures have failed to fully address significant aspects relevant to the viability of physically-orthodox theories-of-*consciousness*, for four reasons. First, discussions of the binding problem often fail to distinguish clearly between *the binding problem for behavioral encoding* (in which spatially distributed information in parallel streams must converge in an orderly yet flexible manner, to support neural computations governing behavior), and *the binding problem for consciousness* (in which spatially-distributed information in parallel streams must be brought together via some as-yet unknown mechanism, to generate the components of conscious experience at each experiential location). Second, even those discussions that disambiguate and then focus on the binding problem for consciousness fail to address *the relationship between spaces*, namely the orthodox physical space (*i.e.* the unseen but putatively existent space in which elementary particles exist) and conscious-experiential space (*e.g.* the visually experienced "emptiness" between distinct objects). Discussion of these two spaces is essential to questions of local causation in the generation of conscious experience, because the two spaces might themselves be topologically unconnected in certain theories-of-consciousness (**Figure 3**), and thus, in a sense to be examined more closely in Section 6, non-local to each other in a basic manner. Third, the question of *temporal locality* in physical causation has been completely neglected, in the context of explaining the neural bases of consciousness. Just as spatial locality becomes questionable when codes are extended across a number of spatial locations, temporal locality must be explicitly accounted for when codes are extended across time in some manner. Fourth, examinations of the binding problem (and consciousness more generally) are usually *couched in informal, verbal, terms*, and thus lack the formal rigor necessary to properly address the behavioral/consciousness binding-problem distinction, the physical-space/conscious-experiential-space relationship, or even a sufficiently complete definition of spatiotemporal locality.

Discussions thus far indicate that a complete answer to questions of locality in the generation of consciousness will need to address all possible physically-orthodox theories-of-consciousness, to consider both distributed and parallel processing, and to remedy the noted deficits in prior approaches. These demands are considerable, and evoke a formal symbolism (summarized in Table 1) for describing relevant aspects of environment, brain, and experience, that will bestow benefits of both generality and precision. Foundations for this formalism are constructed in Section 2 via orienting definitions of salient consciousness-related phenomena. Sections 3 and 4 develop formal symbolism for describing, respectively, the external physical *environment* and consequent brain-dynamical *encodings* of environmental features. Section 5 develops analogous formalism for *conscious experience* of the environment, notably including a coordinate system for relative spatial location of the components-of-consciousness. This coordinate system is then used in Section 6 to describe relevant scenarios for the physical-space/conscious-experiential-space relationship mentioned above, thus providing an informal introduction to locality issues. The relatively lengthy investment in formalism made in Sections 2-5 yields benefits in Section 7, via succinct formal demonstrations that spatially-distributed and temporally-extended codes cannot be the final informational basis for consciousness in a physically orthodox setting. Section 8 describes how the exclusion of these codes constitutes an empirical test for physically-orthodox theories-of-consciousness, and discusses the theoretical implications of alternative empirical outcomes.

2. Consciousness

2.1. Definition

"Consciousness" can be defined *e.g.* as the collection of phenomena that are subjectively present in the awake state of a typical human being, and that are in contrast absent during deep sleep. Although an appeal to subjectivity appears problematic for scientific enquiry that purports objectivity, the subjective/objective distinction must be scientifically acceptable: it is foundationally necessary for the existence of science itself, precisely in order to delineate objective scientific constructions from subjective quotidian experience (for example, to delineate the objective entities constituting the physical environment from the subjective human experience of that environment). Serious problems in the scientific consideration of consciousness *do* arise, however (Hawking, 2000), when theories hypothesizing necessary and sufficient physical conditions for the generation of subjective experience are stated, because there is currently no experimental method for establishing whether an allegedly conscious artefact (Penrose, 1989; Tononi, 2004) does in fact possess subjective experience of any kind. In the case of *human* consciousness, this problem can be partially ameliorated by the presumption that each "typical" human brain has a (currently unknown) mechanism for generating conscious experience, just as each typical human brain has mechanisms *e.g.* for processing afferent sensory data in order to navigate the environment in biologically advantageous ways. In principle, discovery of biophysical mechanisms critically involved in the generation of consciousness (but not in other biological functions) might be experimentally verified in the future by techniques that temporarily interfere with these mechanisms, which should result in disruption of associated components of consciousness (but not of behavior).

The present paper avoids absence-of-experimental-technology problems associated with *general* theories-of-consciousness, because it does not attempt to identify a unique, universally applicable, statement of the physics of conscious experience. Instead, it accepts consciousness as a phenomenon associated with the *human* brain (without any assessment at all of the conscious state of any other biological or non-biological system), and then derives requirements for biophysical mechanisms that must apply if theories-of-consciousness are to be of an entirely physically-orthodox kind. This approach differs radically from most science-of-consciousness papers, whose goal is to establish particular theory-of-consciousness: here, the goal is to establish empirical tests that delineate the entire class of physically-orthodox, mainstream theories [5.2] from those theories requiring either new physics or the unorthodox interpretation of orthodox physics. Thus, the present approach does not depend on experimental tests for conscious *vs.* non-conscious systems, because theoretical developments here do not make predictions concerning this distinction. Nevertheless, the approach here is definitively scientific, rather than philosophical, because it leads to concrete empirical determinations that can arbitrate between two significant theory-classes.

2.2. Limitation to exteroceptive consciousness

Even with a common orienting definition [2.1], the terms “consciousness” and “conscious” are used with a variety of different connotations, leading to the possibility of semantic misunderstanding. Notably, “consciousness” is sometimes used to refer to the controversial idea of a causally-efficacious subject (in statements such as “the ‘conscious self’ made this choice”), whereas present considerations will explicitly *exclude* the conscious self, and instead focus on *exteroceptive* consciousness, *i.e.* on experience of the external environment. That is, without any axiomatic assumptions concerning the objective existence (Popper and Eccles, 1977; Hameroff, 2012) or otherwise of the ‘self’, considerations here will be limited to the subset of consciousness-related phenomena whose biophysical origination depends critically on sensory processing of objective physical features in the external environment. Specifically, developments will be limited to visual and auditory modalities (**Figure 4**), because this limitation of scope offers greatest expository clarity, without loss of generality in results.

2.3. Axiomatic role for brain dynamics in providing information to consciousness

“Psychophysical parallelism” proposes that correspondence between exteroceptive conscious experience and the objective contents of the external environment need not depend on brain dynamics. However, axiomatically in the present work, it will be assumed that brain dynamics are the physical mechanism by which information concerning the environment is transferred to conscious experience. Thus, brain dynamics will be variously described as “constituting the informational basis for consciousness” or “generating a contribution to [or component of] consciousness”, language which explicitly acknowledges a causal role for brain dynamics without proposing any specific theory-of-consciousness. Avoiding adherence to particular theories-of-consciousness is required by the basic approach here of comparing and contrasting two large *classes* of theory [2.1].

3. Formal symbolism for features of the exteroceptive environment

3.1. $s_{a,b}$ notation for perceptual feature types and instances

The brain is thought to parse the external environment with respect to a small finite number of perceptual feature types for each modality (Kandel *et al*., 2012; Zaidi *et al*., 2013). For example, feature types associated with vision include color, motion, oriented edges, and so forth. For the auditory modality, encoding can be conceived in terms of a single feature type, *i.e.* amplitude, indexed secondarily by frequency. (That is, sound at a particular location consists of a set of amplitudes across a variety of frequencies). For each feature type, there are then a number of different specific instances that can be encoded by the brain. For example, for the color feature *type*, the brain is thought to be capable of distinguishing around 10^7 different color *instances* (Judd and Wyszecki, 1975). For expository simplicity, it will be assumed that each feature type has only a finite number of discrete instances, although in principle a continuous code can generate an infinite number of instances. (Explicitly accommodating the continuous code possibility would considerably complicate symbolism, but not affect results.)

The notation $s_{a,b}$ will be adopted for the *b*-th instance of the *a*-th feature type (**Figure 5**). As a rudimentary illustrative example, if the first feature type s_1 is color in the visual field, then $s_{1,1}$ might denote red, $s_{1,2}$ orange, and so on. (More rigorously, $s_{1,1}$ and $s_{1,2}$ should be defined in terms of wavelengths of reflected light, in part because terms such as "red" refer ambiguously both to objective wavelengths and subjective experience: present notation aims to explicitly separate the objective and the subjective, so that the relationship between them can be treated clearly.) The adoption of the basic *s* notation is meant to reinforce that $s_{a,b}$ symbols label various *stimuli* in the exteroceptive environment.

Emphatically, $s_{a,b}$ symbols label *objective* stimulus features that can in principle be reduced to configurations of atoms and molecules, and thence to even more elementary physical constructs (quarks, electrons, and so on). For example, the perceptual feature type of visual color is a way of describing wavelengths of reflected photons, which could equally well be described in terms of the emission and absorption properties of atoms and molecules.

3.2. Reserved usage of {…} notation

Parenthetic expressions of the form "{…}" (as opposed to "(…)" and "[…]") in the present paper explicitly connote multi-member sets or collections. Notably, later symbolism $A(\mathbf{r}_i)$ means the value of *A* at some *particular* coordinate location $\mathbf{r}_i$, whereas $\{A(\mathbf{r}_i)\}$ means the collection of *A*-values at *every* contextually-relevant $\mathbf{r}_i$. Similarly, in present notation, a function $f\,[A(\mathbf{r}_i)]$ depends only on the value of *A* at a single location, whereas a function $g\,[\{A(\mathbf{r}_i)\}]$ depends on *A* at multiple locations: this distinction between single- and multi-location dependence of physical action (or computation) will be pivotal in spatial-locality considerations.

3.3. ${}_{X}\mathbf{r}_j$ notation for locations in the exteroceptive environment

Of course, navigation of the environment by the human organism requires not only identification of perceptual features labelled by $\{s_{a,b}\}$, but also of the locations of these features relative to current position. A complete formal description of the exteroceptive milieu for present purposes therefore requires the assignment of coordinate indices to perceptual features. Because sensory organs (*e.g.* the retina) consist of a finite number of cells with discrete receptive fields, it will be assumed that there are a finite number of fixed environmental locations sampled in exteroception. (Various other assumptions are possible, but do not affect results.) Adopting a conventional notation $\mathbf{r}$ to denote coordinates with respect to some well-defined (*e.g.* brain-centred, Euclidean) spatial axes, the coordinates of exteroceptively relevant locations will be denoted by ${}_{\mathrm{X}}\mathbf{r}_j$ (**Figure 6**), where $j = 1, \ldots, N_E$, and N_E is the total number of locations. (The prefixed subscript "X" labels these coordinates as "eXternal" locations at which various stimulus features $s_{a,b}$ exist, as opposed to later identifiers "A" [4.4] and "B" [5.5.3] that denote two sets of internal brain locations at which dynamical activity has encoding significance.) Thus, the set $\{{}_{\mathrm{X}}\mathbf{r}_j\}$ comprises the coordinates of all locations in the exteroceptive environment from which sensory information originates.

3.4. Completeness and generality of $\{s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j)\}$ notation

3.4.1. Completeness

By construction, an explicit definition of the totality of $s_{a,b}$ at each and every ${}_{\mathrm{X}}\mathbf{r}_j$ constitutes a complete description of the exteroceptive environment, in the sense that neural responses in sensory organs are fully entailed by this information (together with complete understanding of the human organism's neural systems, and full knowledge of the non-sensory states of the organism, such as wakefulness). Note that complete description of the visual environment typically requires the specification of many $s_{a,b}$ at each non-empty ${}_{\mathrm{X}}\mathbf{r}_j$, because, minimally, geometry (*e.g.* orientation of edges), color, and motion must all be specified, and these correspond to three different feature or stimulus types.

3.4.2. Generality

The collection $\{s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j)\}$ is simply a way of describing the external environment with respect to particular set of categories that map in a very direct way to neural encoding (because symbols were defined with this goal in mind). Note that symbols are theory-of-encoding independent, in the following sense. Despite the fact that symbols were developed above with reference to well-established roles for specific features in neural encoding (such as edges, color, and motion, for vision), symbolically-stated theories in the present paper do not rely on the ultimate validity of these roles. In the limit, $\{s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j)\}$ symbols can be taken to stand for the totality of elementary physical constituents of the exteroceptive environment, thus liberating analyses from dependence on any particular theory of encoding, although this approach considerably complicates accompanying verbal exposition (and will not be adopted here for that reason).

4. Formal symbolism for neural encoding of the exteroceptive environment

4.1. Theories of neural encoding

For the purposes of explaining behavior, neural dynamics in sensory organs can be conceived of as encoding the environment (Perkel and Bullock, 1968; Field, 1994; Rieke *et al.*, 1997), in the sense that perfect knowledge of sensory physiology can be used to make certain inferences concerning environmental contents, based on observations of neural dynamics. The relationship between sensory encoding and later motor activity can further be conceived of as a series of computational transformations of the initial encoding (Phillips *et al.*, 1984; Churchland and Sejnowski, 1992), although rigid analogies with human-fabricated electronic computers must be moderated *e.g.* by the role in the brain of contextual dynamical oscillations (Buzsáki, 2006) that have no direct computer analogue. Moreover, despite tremendous advances in the understanding of various types of encoding and computation in sensory and later neural processes [*e.g.* (Olshausen, 1996; Bell and Sejnowski, 1996; Rodriguez *et al.*, 1999; Abbott and Sejnowski, 1999; Mehta *et al.*, 2002; Haynes and Rees, 2006; Quiroga and Panzeri, 2009; Panzeri *et al.*, 2010; Horikawa *et al.*, 2013)], there are many outstanding areas of significant uncertainty (Eggermont, 1998; Van Vreeswijk, 2004) concerning, for example, fidelity of codes in the presence of "noisy" stochastic behavior (Mainen and Sejnowski, 1995; Faisal *et al.*, 2008), the basic organizational unit of encoding (Abeles, 2011), the binding of information encoded in multiple cortical areas (Feldman, 2013), and the encoding role of spatially-distributed electromagnetic-field oscillations (Buzsáki, 2006; Sejnowski and Paulsen, 2006).

In summary, although there is little doubt that the neural coding description is a helpful picture of brain function, currently there is no complete understanding of how information is represented and transformed in neural processing. Thus, any description of how brain encoding contributes information to conscious experience [2.3] must at present be developed in a relatively generic way, specifically via statements that avoid stating a *particular* theory of encoding.

4.2. Notation *A* for brain-dynamical encoding activity

Neural encoding of the exteroceptive environment means that certain brain-dynamical states, but not others, contain the information that a particular instance of a particular perceptual feature type is present at a particular environmental location. In order to formally state neural coding theories, it is necessary to specify a physical *measure* that can distinguish encoding and non-encoding states. The symbol *A* will be used to denote this measure.

4.3. What does *A* stand for?

Three empirical facts support the frequent adoption of an electromagnetic field ("e.m.-field") measure as means of defining and delineating dynamical states relevant to behavioral encoding. First, initial sensory encoding takes place via disturbance of membrane potentials. Second, propagation of information is primarily via electrochemical (axon-to-dendrite-to-axon) and electrical (gap junction) mechanisms. Third, behavioral computations are ultimately converted to motor signals of an electrochemical form. Naturally, it is possible to restate the electrical state of the brain in chemical or biochemical terms, for example, characterizing a "firing" neuron by ion flows or by channel states, rather than by membrane potential. However, for present purposes, these different descriptions are ultimately equivalent, so it will be assumed for discursive simplicity that the behavioral encoding state of the brain is defined by e.m.-field dynamics. Thus, A can be thought of as standing for some measure of e.m.-field state, such as membrane potential.

4.4. $_{\mathrm{A}}\mathbf{r}_i$ notation for encoding-relevant brain locations

To understand what a brain is encoding, one must know the value of A at a set of locations whose coordinates will be denoted by the generic symbol $_{\mathrm{A}}\mathbf{r}_i$ (**Figure 7**). (That is, the set $\{_{\mathrm{A}}\mathbf{r}_i\}$ comprises the coordinates of all brain locations at which A must be known in order to deduce encoded features of the exteroceptive environment. As mentioned in [3.3], the prefixed "A" subscript distinguishes *e.g.* $_{\mathrm{A}}\mathbf{r}_1$ from $_{\mathrm{X}}\mathbf{r}_1$: the former indexes a brain-encoding site, the latter an external environmental location.) The precise loci labeled by the $\{_{\mathrm{A}}\mathbf{r}_i\}$ depend on how the human brain actually encodes exteroception [4.1], but this affects neither their generic definition, nor present developments. For example, the collection $\{_{\mathrm{A}}\mathbf{r}_i\}$ can index *every* brain location.

4.5. C_{abj} notation for activity engendered by $s_{a,b}(_{\mathrm{X}}\mathbf{r}_j)$

By construction, the collection of A measurements at the locations $\{_{\mathrm{A}}\mathbf{r}_i\}$ are sufficient to define the encoding state of the brain. Given a specific stimulus $s_{1,2}$ and a specific environmental location $_{\mathrm{X}}\mathbf{r}_3$, say, it is therefore possible to parse the information $\{A(_{\mathrm{A}}\mathbf{r}_i)\}$ to determine whether $s_{1,2}(_{\mathrm{X}}\mathbf{r}_3)$ is encoded. The precise details of parsing depend again on how the brain actually encodes the environment [4.1], but any reliable encoding method must create a partition between those A-measured brain states that encode a particular $s_{a,b}(_{\mathrm{X}}\mathbf{r}_j)$ (in the general case) and those that do not. This fact can be formally encapsulated by introducing a classifier function C_{abj} on the collection $\{A(_{\mathrm{A}}\mathbf{r}_i)\}$, defined via

$$C_{abj}\,[\{A(_{\mathrm{A}}\mathbf{r}_i)\}] = 1 \;\Leftrightarrow\; \{A(_{\mathrm{A}}\mathbf{r}_i)\} \text{ encodes } s_{a,b}(_{\mathrm{X}}\mathbf{r}_j) \tag{1}$$

and

$$C_{abj}\,[\{A(_{\mathrm{A}}\mathbf{r}_i)\}] = 0 \Leftrightarrow \{A(_{\mathrm{A}}\mathbf{r}_i)\} \text{ does not encode } s_{a,b}(_{\mathrm{X}}\mathbf{r}_j) \tag{2}$$

(where $\Leftrightarrow$ means "if and only if"). The function C_{abj} is then a generic formal placeholder for any specific theory of encoding. For example, if $s_{a,b}(_{\mathrm{X}}\mathbf{r}_j)$ is encoded by the membrane potential of a single neuron, then C_{abj} simply corresponds to the assessment of potential relative to threshold [4.6]. On the other hand, in a complex spatiotemporal encoding, C_{abj} corresponds to determining the presence or otherwise of the particular distributed pattern (or patterns) that encode $s_{a,b}(_{\mathrm{X}}\mathbf{r}_j)$. (Note that, as a default, the argument of $C_{abj}[\ldots]$ is defined as the *multi*-location collection $\{A(_{\mathrm{A}}\mathbf{r}_i)\}$ of A-values [3.2].)

Let the symbols $\exists_{\mathrm{P}}$ and $\nexists_{\mathrm{P}}$ respectively denote the physical (subscript "P") existence and non-existence of an object in a particular setting. Specifically, $\exists_{\mathrm{P}}\, s_{a,b}(_{\mathrm{X}}\mathbf{r}_j)$ will mean that the stimulus $s_{a,b}$ exists in the physical environment at $_{\mathrm{X}}\mathbf{r}_j$, whereas $\nexists_{\mathrm{P}}\, s_{a,b}(_{\mathrm{X}}\mathbf{r}_j)$ indicates that $s_{a,b}$ does not exist at that location. Then, if encoding is a faithful representation of the environment, it follows that

$$\exists_{\mathrm{P}}\, s_{a,b}(_{\mathrm{X}}\mathbf{r}_j) \;\Rightarrow C_{abj}\,[\{A(_{\mathrm{A}}\mathbf{r}_i)\}] = 1 \tag{3}$$

and

$$\nexists_{P}\, s_{a,b}({}_{X}\mathbf{r}_j) \Rightarrow C_{abj}\,[\{A({}_{A}\mathbf{r}_i)\}] = 0 \qquad (4)$$

(where $\Rightarrow$ denotes "… implies that …").

4.6. Example: C_{abj} notation for a single neuron, membrane potential, encoding scheme

A concrete example may help to clarify notation introduced in [4.5]. Consider the stimulus $s_{1,2}$ at the location ${}_{X}\mathbf{r}_3$, and a hypothetical neural encoding in which the presence or absence of $s_{1,2}({}_{X}\mathbf{r}_3)$ leads respectively to firing or non-firing of a neuron whose axon hillock is labelled by coordinates ${}_{A}\mathbf{r}_4$. If A_{thresh} is the minimal membrane potential that is always and only exceeded during firing, then the stimulus-firing relationship can be formally stated as

$$\exists_{P}\, s_{1,2}({}_{X}\mathbf{r}_3) \;\Rightarrow A({}_{A}\mathbf{r}_4) > A_{\text{thresh}} \qquad (5)$$

and

$$\nexists_{P}\, s_{1,2}({}_{X}\mathbf{r}_3) \;\Rightarrow A({}_{A}\mathbf{r}_4) \le A_{\text{thresh}} \qquad (6).$$

Eqs. 5-6 simply state the physical and biophysical logic that presence or absence of stimulus leads to firing or non-firing. Recalling that C_{123} is defined (Eqs. 1-2) with respect to whether $\{A({}_{A}\mathbf{r}_i)\}$ encodes $s_{1,2}({}_{X}\mathbf{r}_3)$, it follows that

$$C_{123}[\{A({}_{A}\mathbf{r}_i)\}] = 1 \;\Leftrightarrow\; A({}_{A}\mathbf{r}_4) > A_{\text{thresh}} \qquad (7)$$

and

$$C_{123}[\{A({}_{A}\mathbf{r}_i)\}] = 0 \;\Leftrightarrow\; A({}_{A}\mathbf{r}_4) \le A_{\text{thresh}} \qquad (8).$$

Note that, although C_{123} on the left-hand side ("LHS") of Eqs. 7-8 takes the complete set of values $\{A({}_{A}\mathbf{r}_i)\}$ as its argument, its one-or-zero value is set only with reference to $A({}_{A}\mathbf{r}_4)$. Thus, the large set of all possible combinatorial instances of $\{A({}_{A}\mathbf{r}_i)\}$ is partitioned into the subset in which $A({}_{A}\mathbf{r}_4) > A_{\text{thresh}}$ (independent of A-values at other ${}_{A}\mathbf{r}_i$) and the disjoint subset in which $A({}_{A}\mathbf{r}_4) \le A_{\text{thresh}}$ (again independently of other A-values).

C_{123} can then be used to replace the $A({}_{A}\mathbf{r}_4)$ firing/non-firing inequalities on the right-hand side ("RHS") of Eqs. 5-6, to give

$$\exists_{P}\, s_{1,2}({}_{X}\mathbf{r}_3) \;\Rightarrow C_{123}[\{A({}_{A}\mathbf{r}_i)\}] = 1 \qquad (9)$$

and

$$\nexists_{P}\, s_{1,2}({}_{X}\mathbf{r}_3) \;\Rightarrow C_{123}[\{A({}_{A}\mathbf{r}_i)\}] = 0 \qquad (10).$$

For this particular encoding, Eqs. 9 and 10 are then directly the $a = 1$, $b = 2$, $j = 3$ instances of the C_{abj} relationships in Eqs. 3 and 4.

4.7. Notation capable of describing any encoding scheme

Eqs. 3-4 are, by construction, simply very general formal statements of the basic premise of any theory of neural encoding [4.1], namely that the presence or absence of a particular stimulus in the environment leads to distinguishable patterns of brain activity. Thus, notation is clearly capable of describing any neural encoding scheme.

5. Formal symbolism for conscious experience of the exteroceptive environment

5.1. Approach to theories-of-consciousness

As noted in [2.1], the present paper does not aim to state or prove a particular theory-of-consciousness. Rather, it discusses the consequences of locality in physical causation for *all* theories of consciousness (after the exclusion of psychophysical parallelism [2.3]). In particular, results here will distinguish between biophysical encoding requirements for “physically orthodox” theories of consciousness [5.2], and for the complementary collection of theories that do not meet all the defining requirements of physical orthodoxy.

5.2. “Physically orthodox” theories-of-consciousness

As noted in the Introduction, the current mainstream assumption in physics and neuroscience is that consciousness can be explained within a physically-orthodox setting, either as synonymous with, or as an emergent property of, brain dynamics. Notably, the mainstream view denies the need for “new physics” of any kind in the explanation of consciousness. A “physically-orthodox setting” is taken here to *minimally* possess three specific properties, namely that physical causation is local, that there are only three spatial dimensions, and that physical structures are composed only of Standard Model elementary particles, such as quarks, electrons, and photons. Although the second and third assumptions are undoubtedly implicit in every mainstream theory-of-consciousness, their explicit emphasis might appear somewhat arbitrary, or even peculiar. However, as described in [7.2.6], these two uncontroversial assumptions (in mainstream settings) create significant limits on *how* the contents of conscious experience can be physically constituted relative to brain dynamics, by ruling out intermediary connective structure that might in principle assimilate distributed codes, and thus overcome locality problems.

5.3. $<s_{a,b}>$ notation for conscious experience of perceptual feature types and instances

Formal symbolism thus far can describe the external environment [via symbols $s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j)$], brain activity [via symbols $A({}_{\mathrm{A}}\mathbf{r}_i)$], and the relationship between environment and brain activity [via symbols C_{abj}]. However, no symbolism has yet been declared for *conscious experience* of the environment. The definition of $s_{a,b}$ symbols allows a very direct construction of required symbolism for consciousness. Introducing notation < ... > to denote “the conscious experience that results from the existence of ... in the exteroceptive environment”, the total conscious exteroceptive experience can be written as $<\{s_{a,b}({}_{X}\mathbf{r}_j)\}>$. Similarly, the contribution to conscious experience from a single stimulus $s_{a,b}({}_{X}\mathbf{r}_j)$ is formally labelled as $< s_{a,b}({}_{X}\mathbf{r}_j) >$ (**Figure 8**).

5.4. ρ_k notation for locations within exteroceptive conscious experience

Self-evidently [2.1], exteroceptive experience consists of a variety of stimulus-induced components arranged in an orderly manner in a spatially-extended setting. The generic symbol ρ will be used to denote a location within the space of conscious experience (**Figure 9**), just as the symbol **r** [3.3] is used to index objective space. As a simplification (that does not affect results), it will be assumed that exteroceptive conscious experience can be fully described by specifying the contents of consciousness at a finite number of locations denoted by $\{\rho_k\}$. Specifically, if sensory information is completely defined by specifying the set $\{s_{a,b}\}$ at ${}_{X}\mathbf{r}_i$, where $i = 1, \ldots, N_E$ [3.3], then exteroceptive conscious experience is completely defined by specifying the $\{<s_{a,b}>\}$ at each ρ_k, where $k = 1, \ldots, N_E$ and ρ_k labels the location in conscious-experiential space corresponding to ${}_{X}\mathbf{r}_k$ in objective space.

The introduction of spatial coordinates for conscious experience raises the typically neglected issue of the relationship between objective and subjective spaces (Dennett and Kinsbourne, 1992). For example, consider the “space” comprised of the emptiness between and surrounding various well-defined objects in visual conscious experience. What, if any, is the correlate in objective reality of experiential space? This complex question cannot be answered in any scientifically definitive way at present. (For example, unobservable physically-orthodox “space” only has a concrete existence in a conventionally realist ontology, whose unique validity cannot currently be established.) For present purposes it is only necessary to establish two points. First, the attribution of coordinates to locations in conscious experiential space is scientifically valid [5.4.1]. Second, the introduction of the generic and specific symbols ρ and ρ_k is entirely general (*i.e.* does not introduce any particular theory-of-reality or theory-of-consciousness) [5.4.2].

5.4.1. Scientific validity of coordinates for locations in conscious experiential space

Although not typically considered explicitly by computational neuroscience, ρ coordinates (or their informal proxies) have been long studied within visual psychophysics (Luneberg, 1944; Foley, 1978; Heller, 1997), where one significant focus has been the characterization of visual as opposed to objective geometry (Wagner, 2006). Consider, for example, the subjective visual experience of the sky on a clear night, in which far distant stars are apparently embedded in a spherical dome, contrasting with the approximately flat spatial geometry understood (*e.g.* by relativistic cosmology) to pertain objectively. This simple comparison shows that subjective and objective geometries need not coincide, and has led to extensive and sophisticated investigations of the general relationship between the two [*e.g.* (Foley *et al.*, 2004)]. In this context, the ρ coordinate system is the basic labelling nomenclature for relative locations within subjective geometry, whereas **r** coordinates are the basis of objective geometry.

5.4.2. Generality of $\{\mathbf{r}, \rho\}$ formal symbolism

The fact of two geometries [5.4.1] does not at all *imply* physical duality of spaces, *i.e.* that locations labeled by $\boldsymbol{\rho}$ coordinates and those labeled by $\mathbf{r}$ coordinates constitute two ontologically equivalent but physically unconnected domains (although neither is this possibility *excluded*, in the general case). For example, the space comprised of locations labeled by $\boldsymbol{\rho}$ coordinates might, as a whole, be a *property of* objective dynamics occurring in the space of locations labeled by $\mathbf{r}$ coordinates, so that the two spaces are physically and ontologically incommensurable. Another possibility is that objects with $\boldsymbol{\rho}$ coordinates and those with $\mathbf{r}$ coordinates might exist in the *same* physical space, so that $\boldsymbol{\rho}$ labels are just different names for points already conventionally labeled by $\mathbf{r}$ coordinates, thus completely eliminating two-ness of constructs. The point for present purposes is that these possibilities (and others) are simultaneously accommodated by the scientifically valid [5.4.1] introduction of $\boldsymbol{\rho}$ coordinates *without* accompanying reduction in current uncertainties concerning the physical, ontological and topological aspects of the $\boldsymbol{\rho}$-$\mathbf{r}$ relationship (other than the required establishment of a *representational* relationship between $\boldsymbol{\rho}_k$ and $\mathbf{r}_k$, namely that $\{s_{a,b}\}$ stimulus features at $\mathbf{r}_k$ are represented by $\{<s_{a,b}>\}$ experiences at $\boldsymbol{\rho}_k$).

5.5. *B* notation for final brain-dynamical cause of contribution to conscious experience

5.5.1. Naïve theory-of-consciousness employing $C_{abj}[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}]$

It is perhaps tempting to move directly from notation developed so far to a formal statement of a general theory-of-consciousness, as follows. Because the presence of $s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j)$ in the exteroceptive environment causes brain dynamics satisfying $C_{abj}[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}] = 1$, it seems reasonable to propose that $C_{abj}[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}] = 1$ is a necessary and sufficient condition for the existence of $<s_{a,b}>(\boldsymbol{\rho}_j)$ (where $<s_{a,b}>(\boldsymbol{\rho}_j)$ denotes the existence of $<s_{a,b}>$ at $\boldsymbol{\rho}_j$ [5.4]), *i.e.*

$$C_{abj}[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}] = 1 \Leftrightarrow \exists_{\mathrm{P}} <s_{a,b}>(\boldsymbol{\rho}_j) \qquad (11)$$

and

$$C_{abj}[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}] = 0 \Leftrightarrow \nexists_{\mathrm{P}} <s_{a,b}>(\boldsymbol{\rho}_j) \qquad (12).$$

Notably, because existence of $s_{a,b}$ at ${}_{\mathrm{X}}\mathbf{r}_j$ implies that $C_{abj}[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}] = 1$ (Eq. 3), Eq. 11 implies, as expected, that

$$\exists_{\mathrm{P}}\, s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j) \Rightarrow \exists_{\mathrm{P}} <s_{a,b}>(\boldsymbol{\rho}_j) \qquad (13)$$

i.e. the existence of a stimulus in the exteroceptive environment leads to the corresponding conscious experience at the appropriate location in subjective space.

5.5.2. Biomolecule basis-of-consciousness example: failure of $C_{abj}[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}]$ theories

Although Eqs. 11-12 are a significant advance (offering the first formal expression of a relationship between brain dynamics and components of conscious experience), they overstate causality and fail to provide an entirely general theory-of-consciousness, as demonstrated by the following example.

Consider a hypothetical situation in which the contribution-to-consciousness from the dynamical activity of a particular neuron is governed, as a fact of nature, by the state of a specific biomolecule M, rather than by the values of the e.m.-field measure A. Because A is presumed to be the general carrier of information [4.3], the consciousness-relevant aspect of M must couple physically with A in a reasonably close way, so that *e.g.* neuronal firing always puts M into the state (or states) that generate the appropriate component of conscious experience. For example, if M has a non-uniform charge distribution (so that e.m-field disturbances from firing can affect M-state), firing can cause a particular, hypothetically consciousness-relevant, energy state of M. The point is that, in this example, it is some aspect of M that is the final physical cause of contribution-to-consciousness, not simply the A-state of the neuron.

When applied to this example, Eqs. 11-12 *overstate causality*, because $C_{abj}[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}] = 1$ is not sufficient to ensure the generation of the conscious experience $<s_{a,b}>(\boldsymbol{\rho}_j)$: if the M molecule is missing or defective, the neuron can be in an encoding A-state but no contribution to consciousness will occur. Formalism thus far is also *an incomplete account of causality* for this example, because there are no formal symbols yet for the causal role of M.

5.5.3. Definition of B and ${}_{\mathrm{B}}\mathbf{r}_i$

The precise final physical cause of brain-dynamical contribution-to-consciousness is not presently known: it could be the A-state of a neuron, the state of some dedicated A-coupled biomolecule M [5.5.2], or any of a number of other physical features that are coupled to or accompany A-dynamics (*e.g.* ion flows or concentrations, vesicle exocytosis, channel states, and so on.) Thus, any general formalism for theories-of-consciousness must accommodate the possibility of a final physical cause for contribution-to-consciousness other than A. The symbol B will be used denote a measure of this *final physical cause* [with the understanding that, in e.m.-field-based theories-of-consciousness (McFadden, 2002; Pockett, 2002), B and A are two symbols for one physical measure]. For example, if energy of a biomolecule is the final physical cause (as imagined in [5.5.2]), B might be defined in terms of a set of distances relative to fixed, local, cellular structure.

Because the location at which B is measured need not be the same as that at which A is measured, the symbols $\{{}_{\mathrm{B}}\mathbf{r}_i\}$ will be introduced to denote the B-measurement locations. For example, in the biomolecule case [5.5.2], ${}_{\mathrm{A}}\mathbf{r}_i$ might denote the location of the axon hillock, and ${}_{\mathrm{B}}\mathbf{r}_i$ the (nearby) location of the biomolecule M.

5.6. D_{abj} notation for contribution to conscious experience engendered by $s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j)$

Just as C_{abj} classifies those states of A that encode $s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j)$, so the function D_{abj} is defined to classify those B-states that generate $<s_{a,b}>(\boldsymbol{\rho}_j)$. Specifically, as direct analogies to Eqs. 1-2,

$$D_{abj}[\{B({}_{\mathrm{B}}\mathbf{r}_i)\}] = 1 \Leftrightarrow <s_{a,b}>(\boldsymbol{\rho}_j) \text{ exists in conscious experience} \quad (14)$$

and

$$D_{abj}\,[\{B({}_{\mathrm{B}}\mathbf{r}_i)\}] = 0 \;\Leftrightarrow\; <s_{a,b}>(\boldsymbol{\rho}_j) \text{ does not exist in conscious experience} \qquad (15).$$

To complete a theory-of-consciousness (**Figure 10**), a physical coupling [5.5.2] between *A*-states and *B*-states must be declared, *e.g.* via

$$C_{abj}\,[\{A({}_{\mathrm{A}}\mathbf{r}_i)\}] = 1 \Leftrightarrow D_{abj}\,[\{B({}_{\mathrm{B}}\mathbf{r}_i)\}] = 1 \qquad (16).$$

After declaration of Eqs. 14-16, the causal chain from stimulus to conscious experience (Eq. 13) still holds, but now depends on a C_{abj}-to-D_{abj}-to-$<s_{a,b}>(\boldsymbol{\rho}_j)$ path, thus remedying the two deficits noted in [5.5.2] for a purely C_{abj} theory.

5.7. Notation capable of describing any theory-of-consciousness

Notation so far does not limit subsequent developments to any particular theory of consciousness (other than requiring that some *physical B*-measure participate in specification of the contents of consciousness [2.3]). That is, $D_{abj}[\{B({}_{\mathrm{B}}\mathbf{r}_i)\}]$ formalism (Eqs. 14-16) is consistent with a wide range of proposals, notably including all physically-orthodox theories. Further development of notation is certainly required to delineate *formally* between various theoretical sub-classes, for example, to distinguish between identity theories (Place, 1956; Smart, 1959) and emergent-property theories (Van Gulick, 2001). These sub-classes make distinct proposals concerning the precise physical properties that $D_{abj}[\{B({}_{\mathrm{B}}\mathbf{r}_i)\}]$ conditions modulate in order to generate conscious experience. However, formalism for these distinctions will not be developed here, because symbolism for coordinate locations and for causal effects already explicitly present in Eqs. 14-16 fully suffices for locality considerations.

6. The relationship between r-space and ρ-space

Notational developments thus far (summarized in Table 1) have been entirely general. (Although certain defining aspects of physically-orthodox theories-of-consciousness were noted in [5.2], these limitations were not used to constrain subsequent formalism.) In this Section and the next, the approach shifts to consider only physically-orthodox theories, thus dividing theories into two classes. The present Section further divides physically-orthodox theories into two sub-classes, according to their treatment of the relationship between **r**-space and **ρ**-space [5.4]. This sub-division of orthodox theories clarifies the basis for assessing locality.

6.1. Physical orthodoxy admits only two basic conceptions of ρ-space

As noted in [5.4.2], **ρ**-**r** symbolism is a general notation, and can describe both dualist and non-dualist theories-of-consciousness. However, under the limitation to physical orthodoxy [5.2], there can be no physically-dual space. Moreover, a physically orthodox approach rules out higher-dimensional constructions that embed orthodox three-space and conscious-space as non-overlapping (disjoint, orthogonal) subspaces of a larger, topologically-connected space (Rosseinsky, 2014b). This means that only two logical possibilities exist for the relationship between **r**-space and **ρ**-space. In the first possibility (here termed the "sub-domain" approach), ρ-coordinates label a sub-domain of orthodox three-dimensional space that is employed for the generation of consciousness (**Figure 11A,B**). In the second (the "property" approach), the conscious-experiential space itself is a property of brain dynamics (**Figure 11C**), so that ρ-coordinates label a kind of virtual or contingent "mental space". This second case contains two distinct spatial constructs, namely the orthodox three-dimensional space and the space-as-property, and may therefore appear to be a dualist conception antithetical to physical orthodoxy. However, it is possible to construe space-as-property as having an ontologically junior status, thus in a sense preserving the "only three [ontologically fundamental] spatial dimensions" requirement of physical orthodoxy [5.2].

6.2. Locality in sub-domain and property conceptions of ρ-space

Physically-orthodox theories must employ either a sub-domain or a property approach to construct a space *in which* $\langle s_{ab}\rangle(\rho)$ experiences occur. This observation is useful to locality considerations, because the two approaches raise different issues.

In the *sub-domain* approach [6.2.1], $\langle s_{ab}\rangle(\rho)$ experiences and B-dynamics that encode them occur in the same **r**-space, so that assessing locality is the rather straightforward matter of whether $\langle s_{ab}\rangle(\rho)$ and B-dynamics are co-located. (For completeness, the discussion of this point in [6.2.1] will also give an introductory treatment of the slightly less straightforward matter of how $\langle s_{ab}\rangle(\rho)$ components come to be arranged appropriately within conscious experience. This question arises because locality constrains the **r**-location of $\langle s_{ab}\rangle(\rho)$ to be the same as its encoding dynamic, thus arranging the $\langle s_{ab}\rangle$ in **r**-space according to the relative locations of neurons.)

In the *property* approach [6.2.2], arrangement of $\langle s_{ab}\rangle(\rho)$ components within conscious experience is not problematic (because the existence of $\langle s_{ab}\rangle$ at ρ_j is taken to be a property of certain B-states, and "at ρ_j" in a property approach means "at the location indexed by ρ_j within conscious-experiential space *that is independent of orthodox space*"). In property-based theories the key issue is instead the conditions under which an **r**-location and a ρ-location can *ever* be "local" to each other, because the property conception inherently proposes that there is no meaningful distance metric between **r**- and ρ-locations. (If such a metric existed, **r**- and ρ-spaces are embedded in a common metric space. Although this is a possible theoretical proposal, it does not fall into the class of physically orthodox theories because common embedding implies a definite ontological equivalence between **r**- and ρ-spaces [6.1].)

6.2.1. Locality in the sub-domain approach

6.2.1.1. Non-distributed and distributed codes in the sub-domain approach

Let $\mathbf{r}(\rho_j)$ denote the **r** coordinate that can be assigned to any ρ_j in the sub-domain approach (but not the property approach). Then the construction of conscious experience is only physically local if the **r**-coordinate of the B-dynamics generating $\langle s_{ab}\rangle(\rho_j)$ is identical to $\mathbf{r}(\rho_j)$. Clearly, this means that all B-dynamics generating $\langle s_{ab}\rangle(\rho_j)$ must be at the *same* location [namely, $\mathbf{r}(\rho_j)$]. Because distributed codes employ B-dynamics at multiple distinct locations, they cannot be local. Additionally, a non-distributed code must create $\langle s_{ab}\rangle(\rho_j)$ at the same **r**-location as its encoding *B*-dynamic, if the construction of experience is to be local.

6.2.1.2. The construction of subjective geometry in the sub-domain approach

A local sub-domain approach must employ a non-distributed code in which $\langle s_{ab}\rangle(\rho_j)$ is generated at the same **r**-location as its encoding *B*-dynamic [6.2.1.1]. Because the space of conscious experience is orthodox physical space in a sub-domain approach, this means that the collection $\{\langle s_{ab}\rangle(\rho_j)\}$ will be arranged within conscious experience *just as* corresponding *B*-encodings are arranged within orthodox physical space. There are then two possibilities. First, if *B*-encodings are arranged within orthodox physical space in a manner that recapitulates the geometry of the $\{\langle s_{ab}\rangle(\rho_j)\}$ within subjective experience [5.4.1], then (at least for a single feature type) brain dynamics generate conscious experience in a direct and consistent manner. [Additional complications for multiple feature types are discussed in (Rosseinsky, 2014b).] Second, if the **r**-configuration of *B*-encodings does *not* recapitulate the subjective geometry of the $\{\langle s_{ab}\rangle(\rho_j)\}$, some further mechanism must exist to rearrange the $\{\langle s_{ab}\rangle(\rho_j)$. To facilitate this rearrangement, it is possible to propose that space exists under two metrics, the conventional **r**-metric and a **ρ**-metric that puts points close together in conscious experience that are far apart under the conventional metric. [Note that elaboration of this "dual metric" approach is necessary for a rigorously comprehensive treatment of physically orthodox theories-of-consciousness, in the following sense. In the absence of a method for generating subjective geometry from arbitrarily arranged *B*-dynamics, local sub-domain theories can only be given if the **r**-configuration of *B*-dynamics already recapitulates subjective geometry, thus severely limiting the physically-orthodox class. The question of whether dual-metric theories themselves qualify as "physically orthodox" is discussed elsewhere (Rosseinsky, 2014b). For present purposes, the approach will be to establish the *broadest possible* set of physically orthodox theories, and then to show that, within this set, only non-distributed codes are local.]

6.2.2. Locality in the property approach

6.2.2.1. r/ρ-locality

In one view, every property-based theory is *non-local*, because points indexed by **ρ**-coordinates have no meaningful distance from any **r**-location. In this view, the inability to establish locality (*i.e.* a *zero* distance) is non-locality, leading to simplification of the locality-of-physically-orthodox-theories question: only sub-domain theories can be local (and therefore, after [6.2.1.1], the only local physically-orthodox theories are non-distributed codes in which $\langle s_{ab}\rangle(\rho_j)$ is generated at the same **r**-location as its encoding *B*-dynamic).

In an alternate view, every theory employing the property approach is *local*, again because **ρ**-coordinates have no meaningful distance from any **r**-location! In this view, the inability to establish non-locality (*i.e.* a *non*-zero distance) is locality. One notable extension of this view is that a **ρ**-location can then be local (or more precisely "not **r**-non-local") to a multiplicity of **r**-locations. Note that *B*-properties are defined at **r**-locations rather **ρ**-locations. But *if B*-properties could exist "in" **ρ**-space, it would then be possible for a single **ρ**-location to "contain" *B*-values from a number of **r**-

locations, so that a multiple-$\mathbf{r}$ (*i.e.* distributed) code could be local with respect to $\boldsymbol{\rho}$. Allowing this chain of speculative possibilities would then lead to the result that distributed codes *can be* local in certain property approaches, an implication of fundamental significance to central considerations

These discussions establish that the way in which locality is attributed to property approaches (**Figure 3D,E**) will profoundly affect the final view of locality in physically-orthodox theories. Therefore, if final results are to be reliable, an objective way of making this attribution must be established. This can be done in two steps, as follows. First, because the central result will *exclude* distributed codes from local, physically-orthodox, theories-of-consciousness, the attribution should be made in the first instance in a manner that *includes* as many distributed coding possibilities as possible. This means adopting the view that property approaches are naturally local, or more precisely “not $\mathbf{r}$-non-local”, by virtue of the metric absence. Second, the locality (or not-non-locality) thus made available to property approaches must be limited by the overall conception of property approaches as *physically orthodox*. Recalling that concerns of a non-orthodox duality in property approaches are to be answered by an ontologically junior status for conscious-experiential space [6.1], this means that *B*-dynamics cannot exist in $\boldsymbol{\rho}$-space, resulting in a critical constraint on the locality of property approaches: $\mathbf{r}/\boldsymbol{\rho}$-locality in a property approach cannot be used to facilitate or perform joint computations across *B*-values at multiple $\mathbf{r}$-locations.

6.2.2.2. Non-distributed and distributed codes in the property approach

The definition of $\mathbf{r}/\boldsymbol{\rho}$-locality established in [6.2.2.1] has the following implications. First, *non*-distributed codes in the property approach are all local. (Note that in the sub-domain approach, it is possible for a *non*-distributed code to be *non*-local, if $\mathbf{r}(\boldsymbol{\rho}_j)$ is not identical to the ${}_{B}\mathbf{r}$ coordinate of the corresponding *B*-encoding). Second, distributed codes in the property approach are all non-local, because joint *B*-computations across multiple $\mathbf{r}$-locations must be completed at a single point in $\mathbf{r}$-space, before the result of that computation generates some $\langle s_{ab}\rangle(\boldsymbol{\rho}_j)$ (as the *property-of* the completed *B*-computation value).

7. Locality requires that point-like states encode components-of-consciousness

Developments in Section 6 informally establish that locality requires non-distributed codes for consciousness (because [6.2.1.1] established that sub-domain approaches can only be local if codes are non-distributed, and [6.2.2.2] established that all distributed codes in the property approach are non-local). Although $\boldsymbol{\rho}$-formalism [5.4] is foundational to the somewhat intuitive approach employed in Section 6, the full *D* symbolism for theories-of-consciousness [5.6] was not employed there. Nor were the demonstrations of Section 6 premised on a *formal* definition of local causality. The present Section remedies these deficits by providing a formal analysis of locality in physically-orthodox theories. Formal proofs are not entirely independent of the intuitive developments of Section 6, however, because they rely on the sub-domain/property bifurcation established there [6.1], as well as on the definition of $\mathbf{r}/\boldsymbol{\rho}$-locality in [6.2.2.1]. Nevertheless, formal developments in the present Section supply both an additional degree of *rigor*, and an extension in the *generality* of analyses, to consider parallel and temporal codes deliberately neglected in Section 6. This Section also addresses the key question of why physical orthodoxy excludes distributed *B*-coding of consciousness but does *not* exclude distributed *A*-codes for neural computations serving behavior.

7.1. Locality requires that point-like states encode components-of-consciousness

7.1.1. Formal characterization of local causality

Let P_1 be a symbol for the measured value of a physical property at a location with coordinates $\boldsymbol{\chi}_1$, and P_n the analogous symbol for some other property at $\boldsymbol{\chi}_n$. Then "locality of causation" means that a change ΔP_1 at $\boldsymbol{\chi}_1$ can only cause a change ΔP_n at $\boldsymbol{\chi}_n$ if $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_n$ are two labels for a single physical location, or if there is a continuous path $\boldsymbol{\chi}_2, \ldots, \boldsymbol{\chi}_{n-1}$ from $\boldsymbol{\chi}_1$ to $\boldsymbol{\chi}_n$ such that ΔP_1 results in a physical change ΔP_2 at $\boldsymbol{\chi}_2$, ΔP_2 results in ΔP_3 at $\boldsymbol{\chi}_3$, and so on, until ΔP_{n-1} results in the change ΔP_n at $\boldsymbol{\chi}_n$.

For present purposes, a change in neural state $\Delta A({}_{\mathrm{A}}\mathbf{r}_1)$ at some location ${}_{\mathrm{A}}\mathbf{r}_1$ can certainly have distant effects, say on muscle activity (*e.g.* after intermediate, behaviorally-relevant, neural computations), if there is a path of physical connectivity from ${}_{\mathrm{A}}\mathbf{r}_1$ to relevant muscle sites. Any such spatially-distant effect is then implemented via a chain of local causes and effects. Consider, in contrast, the effect of $\Delta B({}_{\mathrm{B}}\mathbf{r}_1)$ on the component of consciousness $<s>$ at $\boldsymbol{\rho}_1$, say. $\Delta B({}_{\mathrm{B}}\mathbf{r}_1)$ can only engender a change $\Delta<s>$ at $\boldsymbol{\rho}_1$ that is *consistent with local causation* if the coordinates ${}_{\mathrm{B}}\mathbf{r}_1$ and $\boldsymbol{\rho}_1$ label the same physical location, or if there is a connected physical path from ${}_{\mathrm{B}}\mathbf{r}_1$ to $\boldsymbol{\rho}_1$ that also contains causally efficacious physical structures, transmitting $\Delta B({}_{\mathrm{B}}\mathbf{r}_1)$ to $\Delta<s>$ via a chain of local changes in physical state.

The analysis of local causality in the generation of conscious experience therefore requires some knowledge of the relative location, ontology, and topology of locations labelled by $\mathbf{r}$ coordinates and those labelled by $\boldsymbol{\rho}$ coordinates. Although, as discussed in [5.4], considerable uncertainty currently exists regarding these matters in the *general* case, comprehensive analysis in Section 6 of possibilities in *physically-orthodox* settings will suffice as a firm foundation for the formal analyses in [7.2-7.3].

7.1.2. Locality in classical and quantum physical theories

Spatiotemporally-local causality is built into modern physical theory, because almost all phenomena amenable to an orthodox explanation must be reducible to the existence and local interactions of elementary particles and spacetime curvature. Although quantum phenomena are sometimes advanced as a broad exception to local causality, for present purposes, it suffices to note that quantum non-locality (if it exists) is not of an arbitrary kind that can be harnessed to any purpose. Moreover, because the central present focus is to characterize a biophysical test for physically-orthodox theories-of-consciousness that are (as noted in the Introduction) all purely *classical* in nature, further discussion of quantum non-locality will be given as a discussion point [8.1.3], rather than as a feature of central analysis.

7.2. Spatial locality

7.2.1. Spatial locality excludes distributed *B*-codes for a single feature type

The intuitive discussion of consequences from spatial locality for a physically orthodox theory-of-consciousness given in [6.2] can be stated in more rigorous formal terms, by employing the $D_{abj}[\{B({}_{\mathrm{B}}\mathbf{r}_i)\}]$ symbolism of Eqs. 14-16. Specifically, consider a single modality (*e.g.* vision) and a distributed code in which activity at multiple locations encodes a single feature type at a single $\boldsymbol{\rho}$ location. For example, let activity at three separate locations ${}_{\mathrm{B}}\mathbf{r}_1$, ${}_{\mathrm{B}}\mathbf{r}_2$, and ${}_{\mathrm{B}}\mathbf{r}_3$ encode the conscious experience of color, $<s_{col,b}>$ say, at $\boldsymbol{\rho}_1$. (In this notation, b indexes the various different colors that can be experienced at $\boldsymbol{\rho}_1$.) Following Eq. 14, writing $D_{\mathrm{col},b,1}$ for the B-classifier relevant to color at $\boldsymbol{\rho}_1$, it follows that

$$D_{\mathrm{col},b,1}[\{B({}_{\mathrm{B}}\mathbf{r}_1), B({}_{\mathrm{B}}\mathbf{r}_2), B({}_{\mathrm{B}}\mathbf{r}_3)\}] = 1 \Leftrightarrow \exists_{\mathrm{P}} <s_{col,b}>(\boldsymbol{\rho}_1) \qquad (17).$$

Locality problems are immediately apparent in Eq. 17, because the LHS requires integration of B-information at multiple locations that are not connected by an $<s_{col,b}>(\boldsymbol{\rho}_1)$-relevant causal physical structure [7.1.1].

7.2.1.1. Case 1: sub-domain scheme

Consider the case in which $\boldsymbol{\rho}$-coordinates are labels for a sub-domain of orthodox physical three-dimensional space [6.1,6.2.1], so that points labeled by ${}_{\mathrm{B}}\mathbf{r}_1$, ${}_{\mathrm{B}}\mathbf{r}_2$, ${}_{\mathrm{B}}\mathbf{r}_3$, and $\boldsymbol{\rho}_1$ can all be viewed as lying in a single, physically-connected, space. Consider further the *degenerate* case in which $\boldsymbol{\rho}_1$ labels a point that is already labeled by one of the $\{{}_{\mathrm{B}}\mathbf{r}_1, {}_{\mathrm{B}}\mathbf{r}_2, {}_{\mathrm{B}}\mathbf{r}_3\}$, ${}_{\mathrm{B}}\mathbf{r}_1$ say (without loss of generality), *i.e.* where the point in orthodox three-dimensional space used to generate $<s_{col,b}>(\boldsymbol{\rho}_1)$ is *the same as* one of the points used for B-measurement of $D_{\mathrm{col},b,1}$ conditions. (The same result follows trivially for the *non-degenerate* or *general* sub-domain case, where $\{{}_{\mathrm{B}}\mathbf{r}_1, {}_{\mathrm{B}}\mathbf{r}_2, {}_{\mathrm{B}}\mathbf{r}_3, \boldsymbol{\rho}_1\}$ all label different points.) In the degenerate case, information must be propagated from ${}_{\mathrm{B}}\mathbf{r}_2$ and ${}_{\mathrm{B}}\mathbf{r}_3$ to $\boldsymbol{\rho}_1 = {}_{\mathrm{B}}\mathbf{r}_1$ in order to assess $D_{\mathrm{col},b,1}$ values, and the lack of intermediate physical structure for this propagation means that the generation of $<s_{col,b}>(\boldsymbol{\rho}_1)$ is *not* local, according to [7.1.1].

Neural connectivity between $\{{}_{\mathrm{B}}\mathbf{r}_1, {}_{\mathrm{B}}\mathbf{r}_2, {}_{\mathrm{B}}\mathbf{r}_3\}$ and $\boldsymbol{\rho}_1$ cannot be the intermediate mechanism for propagating $D_{\mathrm{col},b,1}$ information. For example, consider the general case, and assume that B-measures at $\{{}_{\mathrm{B}}\mathbf{r}_1, {}_{\mathrm{B}}\mathbf{r}_2, {}_{\mathrm{B}}\mathbf{r}_3\}$ were propagated to $\boldsymbol{\rho}_1$ by orthodox physical and neural means (*i.e.* by a chain of physically-orthodox local causes and effects). In this case, the encoding of $<s_{col,b}>(\boldsymbol{\rho}_1)$ is *not distributed*, contradicting the assumption in [7.2.1], because under these circumstances the correct theory of consciousness is

$$D_{\mathrm{col},b,1}[B({}_{\mathrm{B}}\mathbf{r}_4)] = 1 \Leftrightarrow \exists_{\mathrm{P}} <s_{col,b}>(\boldsymbol{\rho}_1) \qquad (18)$$

replacing Eq. 17, where ${}_{\mathrm{B}}\mathbf{r}_4$ is the $\mathbf{r}$-coordinate label for the point also labeled by $\boldsymbol{\rho}_1$. (Eq. 18 must apply under these circumstances, because B was defined [5.5.3] as the final informational basis of consciousness in brain dynamics.)

7.2.1.2. Case 2: property scheme

In a physically orthodox setting, the only alternative to the sub-domain interpretation [6.1,6.2.1] of $\boldsymbol{\rho}$-

space is the property scheme [6.1,6.2.2]. In this conception, points labelled by $\{{}_{B}\mathbf{r}_1, {}_{B}\mathbf{r}_2, {}_{B}\mathbf{r}_3\}$ exist in a common space (the physically-orthodox three-dimensional space), whereas $\boldsymbol{\rho}$-coordinates label points in a separate (physically unconnected) space that is itself the property of brain dynamics. Even if $\mathbf{r}/\boldsymbol{\rho}$-locality is admitted [6.2.2.1], $\{{}_{B}\mathbf{r}_1, {}_{B}\mathbf{r}_2, {}_{B}\mathbf{r}_3\}$-located information must first be integrated to some point ${}_{B}\mathbf{r}_4$, via a definitively $\mathbf{r}$-non-local mechanism, before acting at $\boldsymbol{\rho}_1$ via a putatively local (not-non-local) ${}_{B}\mathbf{r}_4$-$\boldsymbol{\rho}_1$ connection.

7.2.2. Locality does not exclude distributed *A*-coding for behavioral processes

Taken together, [7.2.1.1-7.2.1.2] imply that the informational basis for any particular component-of-consciousness $<s_{a,b}>$ must be spatially point-like, *i.e.* depend only on a *single* ${}_{B}\mathbf{r}$ coordinate. This means that the encoding for consciousness of a particular feature-instance cannot be distributed, and must be contained in a single neuron.

However, these results do *not* contradict *e.g.* the well-established roles for distributed codes in computations that serve the generation of behavior (as opposed to the generation of consciousness). Provided that "wiring" (physical connectivity of an appropriate kind [7.1.1]) links all locations involved in a distributed code with causally-later information processing machinery, local causality is not violated, in these contexts. It is precisely the absence of such wiring in the case of consciousness that leads to the conflict between distributed codes and local causality.

7.2.3. Locality does not exclude parallel *B*-processing of distinct feature types

It is perhaps tempting to conclude from [7.2.1.1-7.2.1.2] that *all* feature types at a single $\boldsymbol{\rho}$-location must be encoded by a single neuron, *i.e.* must depend only on a *single* ${}_{B}\mathbf{r}$ coordinate. This would mean that the generation of conscious experience at a given location $\boldsymbol{\rho}_i$ would have to obey

$$D_W\,[B({}_{B}\mathbf{r}_i)] = 1 \Leftrightarrow \{<s_{a,b}>(\boldsymbol{\rho}_i),\,(a,b)\in W\} \qquad (19)$$

where W is a set of (a,b) pairs defining the totality of features at $\boldsymbol{\rho}_i$, and the member constitution of W is determined solely by the B-measure at a single brain location ${}_{B}\mathbf{r}_i$. (Recall that a one-to-one correspondence between the $\{\boldsymbol{\rho}_i\}$ and the $\{{}_{X}\mathbf{r}_i\}$ was declared in [5.4]. Eq. 19 defines a new one-to-one relationship between the $\{\boldsymbol{\rho}_i\}$ and the $\{{}_{B}\mathbf{r}_i\}$.)

However, reasoning in [7.2.1.1-7.2.1.2] that considers a single feature type could only be extended to all feature types simultaneously if it were known that a single physical computation occurs for each $\boldsymbol{\rho}$-location to determine all features at that location. Consider the alternative, namely that there is a set of computations, one for each feature type. Following [7.2.1.1-7.2.1.2], let ${}_{B}\mathbf{r}_{\mathrm{col},1}$ be the single B-location encoding color at $\boldsymbol{\rho}_1$, and let ${}_{B}\mathbf{r}_{\mathrm{edge},1}$ be the single B-location encoding edge-orientation at $\boldsymbol{\rho}_1$. For simplicity, consider the property scheme [6.1,6.2.2]. Here, it is possible for the equations

$$D_{\mathrm{col},b,1}[B({}_{B}\mathbf{r}_{\mathrm{col},1})] = 1 \Leftrightarrow \exists_{P} <s_{col,b}>(\boldsymbol{\rho}_1) \qquad (20)$$

and

$$D_{\mathrm{edge},c,1}[B({}_{\mathrm{B}}\mathbf{r}_{\mathrm{edge},1})] = 1 \Leftrightarrow \exists_{\mathrm{P}} <s_{edge,c}>(\boldsymbol{\rho}_1) \qquad (21)$$

to pertain, without contradicting locality, if $<s_{col,b}>(\boldsymbol{\rho}_1)$ is interpreted as a property of $B({}_{\mathrm{B}}\mathbf{r}_{\mathrm{col},1})$, and $<s_{edge,c}>(\boldsymbol{\rho}_1)$ as a property of $B({}_{\mathrm{B}}\mathbf{r}_{\mathrm{edge},1})$. (This interpretation relies on the definitional separation [3.1] of $s_{col,b}$ and $s_{edge,c}$, and the consequent fact that there are two separate D equations, *i.e.* Eqs. 20-21 . It cannot be applied to give a local interpretation of a single-feature distributed code such as Eq. 17, because self-evidently this equation inherently contains multi-**r** references.)

In summary, if local causality applies, the a-th feature type must be encoded by a non-distributed code that defines the x-instance $<s_{a,x}>$ for a given $\boldsymbol{\rho}_j$ by the B-value at a *single* **r** location, ${}_{\mathrm{B}}\mathbf{r}_{a,j}$ say. But various different feature types s_a, s_b, s_c, and so on, can have their instances $<s_{a.x}>$ $<s_{b,y}>$ $<s_{c,z}>$ at $\boldsymbol{\rho}_j$ defined by B-values at a *different* **r** locations ${}_{\mathrm{B}}\mathbf{r}_{a,j}$, ${}_{\mathrm{B}}\mathbf{r}_{b,j}$, ${}_{\mathrm{B}}\mathbf{r}_{c,j}$ *etc.* in a manner consistent with local causality, *e.g.* under certain interpretations of the property conception of $\boldsymbol{\rho}$-space.

7.2.4. Locality does not exclude topographic *B*-processing for a single feature type

Emphatically, results in [7.2.1.1-7.2.1.2] do not contradict the feasibility of parallel (as opposed to distributed) processing of a *single* feature in physically-orthodox theories-of-consciousness. Thus, for example, $<s_{1,b(1)}>(\boldsymbol{\rho}_1)$, $<s_{1,b(2)}>(\boldsymbol{\rho}_2)$, $<s_{1,b(3)}>(\boldsymbol{\rho}_3)$ and so on can be encoded as the result of parallel computations that generate B-values at ${}_{\mathrm{B}}\mathbf{r}_{1,1}$, ${}_{\mathrm{B}}\mathbf{r}_{1,2}$, ${}_{\mathrm{B}}\mathbf{r}_{1,3}$ *etc.*, with

$$D_{1,b(j),\,j}\,[B({}_{\mathrm{B}}\mathbf{r}_{1,\,j})] = 1 \Leftrightarrow \exists_{\mathrm{P}} <s_{1,\,b(j)}>(\boldsymbol{\rho}_j) \qquad (22).$$

Eq. 22 can, for example, describe a (non-distributed) *topographic* encoding. Arguments of [7.2.1.1-7.2.1.2] exclude only (and importantly) the appearance of multiple **r**-locations in the argument of D for a given $<s_{a,b}>(\boldsymbol{\rho}_j)$, *i.e.* terms of the form $D_{abj}[\{B({}_{\mathrm{B}}\mathbf{r}_i)\}]$ (recalling from [3.2] that {…} notation in the present work is reserved for multi-member sets or collections).

7.2.5. Spatially-extended phenomena in orthodox physics

One potential confusion concerning results so far must be cleared up. Specifically, it might appear that orthodox physical theory contains spatially-extended properties without intermediate causal structure (**Figure 12**). For example, temperature is a physically orthodox property that appears to pertain to molecules that are spatially distributed, so it might appear that consciousness could similarly be a property of a spatially-distributed code, without violating physical orthodoxy. However, careful analysis of temperature shows that it is a *measurement* involving a *spatially-extended measuring apparatus* (thermometer). Apparatus of this kind constitutes precisely the kind of "intermediate causal structure" [7.2.6] that is absent in a physically-orthodox theory-of-consciousness. Thus analogies between temperature and consciousness are not valid, with respect to spatial extension.

7.2.6. Roles of "3-D space" and "conventional particle spectrum" assumptions

In [5.2], physically-orthodox theories-of-consciousness were specifically restricted to three-dimensional space and a standard particle spectrum. These restrictions implicitly support the arguments of [7.2.1.1-7.2.1.2] by ruling out the *conceptual* possibility of a physical structure beyond the brain that could (even in principle) link encoding sites $\{{}_{B}\mathbf{r}_i\}$ to $\{\boldsymbol{\rho}_j\}$ locations (**Figure 12D**).

If space were *e.g.* six- rather than three-dimensional, one could conceive of {quark, electron, photon}-constituted physical structures linking encoding sites $\{{}_{B}\mathbf{r}_i\}$ in the brain with $\{\boldsymbol{\rho}_j\}$ locations in a three-dimensional subspace orthogonal to that containing the brain. (Extra dimensionality is required if these structures are constituted of orthodox particles, because such structures would be directly observed if they existed in the three-dimensional subspace containing the brain.)

If non-standard particles could bind together to constitute physical structures linking encoding sites $\{{}_{B}\mathbf{r}_i\}$ in the brain with $\{\boldsymbol{\rho}_j\}$ locations, and these structures had no interaction with orthodox matter other than at $\{{}_{B}\mathbf{r}_i\}$ sites, then extra-dimensionality would not be required to house information transmission mechanisms.

Thus, excluding extra-dimensionality and non-orthodox particles (as is conventional, in physically orthodox theories) rules out certain hypothetical possibilities that could contradict the "no intermediate causal structure" claims of [7.2.1.1-7.2.1.2]. Conversely, as will be discussed in [8.1.2], explicitly introducing extra-dimensionality or non-orthodox particles can allow distributed codes to be consistent with local causation, by providing the means for intermediate causal structures.

7.3. Temporal locality

7.3.1. Temporal locality excludes rate and temporal codes

A rate code (Adrian, 1928) necessarily requires the employment of a temporally extended interval over which the rate or frequency of firing is established. More complex temporal codes (Reinagel and Reid, 2000; Mehta *et al*., 2002; Panzeri *et al*., 2010) also require a finite (non-zero) temporal interval for decoding. For present purposes, extended temporal decoding intervals are analogous to the spatial distribution of B-dynamics, in the sense that they create an informational basis for consciousness that is *not* spatiotemporally point-like.

Consider a single neuron rate-code for $<s_{a,b}>(\boldsymbol{\rho}_j)$ in which B activity must be assessed over a temporal interval $\{t, \ldots, t+\varepsilon\}$ (where t is a conventional time coordinate, and $\{t, \ldots, t+\varepsilon\}$ denotes the continuum of points from t to $t+\varepsilon$, including the end points). Writing $B(\mathbf{r},t)$ for the time-t value of B at $\mathbf{r}$, the theory-of-consciousness for such a code must be of the form

$$D_{abj}\,[B({}_{B}\mathbf{r}_i, \{t,\ldots, t+\varepsilon\})] = 1 \Leftrightarrow <s_{a,b}>(\boldsymbol{\rho}_j, \{t+\varepsilon,\ldots, t+\varepsilon+\Delta\}) \qquad (23)$$

where $<s_{a,b}>(\boldsymbol{\rho}_j, t)$ denotes the time-t existence of $<s_{a,b}>(\boldsymbol{\rho}_j)$, and the time interval on the RHS describes the existence of $<s_{a,b}>(\boldsymbol{\rho}_j)$ for a period of length Δ beginning at $t+\varepsilon$. [For example, Δ is order 10^{-1} s, if conscious experience is refreshed every 100 ms or so (VanRullen and Koch, 2003).] The absence of intermediate causal structure to act as a kind of memory, and thus effect temporally extended processing, means that a locally-causal D cannot act on a temporally-extended set of B values in a physically orthodox setting, thus excluding Eq. 23-based theories and a wide range of temporal codes, including rate codes.

Exclusion of spatially-distributed and temporally-extended codes means that a particular feature $<s_{a,b}>$ at a given coordinate location $\boldsymbol{\rho}_j$ must be encoded by the instantaneous value of B at a single point in space and time, for local causality, so that

$$D_{abj}\,[B({}_{\mathrm{B}}\mathbf{r}_i,\, t)] = 1 \Leftrightarrow\, <s_{a,b}>(\boldsymbol{\rho}_j,\, t) \qquad (24).$$

This does not mean that $<s_{a,b}>(\boldsymbol{\rho}_j)$ itself cannot be temporally extended, but only that such temporal extension must be generated by temporally-extended satisfaction of the condition on the LHS of Eq. 24. That is, in order to generate $<s_{a,b}>(\boldsymbol{\rho}_j, \{t_0 ,\ldots, t_0 + \Delta\})$, we must have $D_{abj}[B({}_{\mathrm{B}}\mathbf{r}_i, t)] = 1$ for each t in $\{t_0 ,\ldots, t_0 + \Delta\}$. (Note the difference between this condition and the LHS of Eq. 23. In Eq. 24, the satisfaction of $D_{abj}[B({}_{\mathrm{B}}\mathbf{r}_i, t)] = 1$ for each t has the temporally-local effect that $<s_{a,b}>(\boldsymbol{\rho}_j, t)$ exists. In contrast, in Eq. 23 the action of D requires first the temporally *non*-local accumulation of B-values throughout the interval $\{t,\ldots, t + \varepsilon\}$.)

Eq. 24 challenges a role for spike-based codes as the informational basis of consciousness in a locally-causal theory, because a neuron can only be in the spiking state for a few milliseconds at a time. (That is, even during a temporally-extended high frequency burst, spikes are interspersed with non-spike intervals). Thus, temporally local generation of a component-of-consciousness that itself persists continuously for 100 ms or more cannot be based on spike states, because spike states only persist for a few milliseconds. (At best, if firing were sustained for 100 ms, temporally-local generation must mean that the component-of-consciousness "flickers" in and out of existence in precise synchronization with the spiking/non-spiking state of the neuron.)

Despite the inability of spike-based codes to produce temporally continuous and persistent components of consciousness, other mechanisms are possible. Consider, for example, a neuron that encodes a feature $s_{a,b}({}_{\mathrm{X}}\mathbf{r}_j)$ by spiking activity of some kind, which in turn puts a biomolecule M_{abj} [5.5.2] into some energy state $E[M_{abj}](t_0)$ at time t_0. (In this notation, $E[O]$ is the energy of a physical object labelled by the symbol O, and $E[O](t)$ is the energy of O at time t.) Assume that the classifier D_{abj} for $<s_{a,b}>(\boldsymbol{\rho}_j)$ can be written as

$$E[M_{abj}](t) > E_0 \Leftrightarrow\, <s_{a,b}>(\boldsymbol{\rho}_j,\, t) \qquad (25),$$

i.e. that $<s_{a,b}>(\boldsymbol{\rho}_j)$ exists when the energy of M_{abj} is above some threshold E_0. Then, if the burst-induced energy $E[M_{abj}](t_0)$ is greater than E_0, and if a time interval of duration Δ elapses before dissipative mechanisms result in $E[M_{abj}]$ falling below E_0, it follows from Eq. 25 that $<s_{a,b}>(\boldsymbol{\rho}_j)$ exists continuously for the period $<s_{a,b}>(\boldsymbol{\rho}_j, \{t_0 ,\ldots, t_0 + \Delta\})$.

7.3.2. Temporally-extended phenomena in orthodox physics

Just as certain orthodox physical properties might appear to be spatially extended [7.2.5], the appearance of temporal extension in physically-orthodox settings might seem to invalidate the claim that physically-orthodox codes must be temporally point-like. Consider, for example, the simple temporal persistence of a rigid body (in the absence of conditions that would cause it to move or to fragment). Clearly, although temporal extension is real, it results from a succession of temporally-local persistences of component elementary-particles. In contrast, the conversion of a temporally-extended code (*e.g.* Eq. 23) involved in the generation of some component-of-consciousness requires non-temporally-local *memory* of physical features, meaning that the analogy between temporal extension of the rigid body and temporally-extended computation breaks down.

8. Discussion

8.1. Experimental tests and implications from future experimental results

8.1.1. Biophysical structures must convert distributed to local codes

Results in [7.2] and [7.3] mean that physically-orthodox theories-of-consciousness must propose spatiotemporally point-like encodings of contributions-to-consciousness, if these theories are to be consistent with locality of physical causation. This implication contrasts with the widespread prevalence of spatially-distributed (Georgopoulos *et al*., 1986; Buonomano and Merzenich, 1995; Tsodyks *et al*., 1996; Pillow *et al*., 2008; Solomon and Lennie, 2007; Quiroga and Panzeri, 2009) and temporally extended (Adrian, 1928; Reinagel and Reid, 2000; Mehta *et al*., 2002; Panzeri *et al*., 2010) encoding schemes in current accounts of neural information processing that serves behavior. Point-like encoding of consciousness and distributed/extended behavioral codes can only be reconciled if biophysical mechanisms exist in the brain to convert non-point-like codes to spatiotemporally local equivalents, for each encoding that is used as the informational basis of consciousness. These mechanisms must be empirically observable, thus creating a strong experimental constraint for locally-causal generation of contributions-to-consciousness based on brain dynamical states. Potential methods for verification or falsification of these predictions include detailed neuroanatomical investigations of dynamically-relevant brain structure (Alivisatos *et al*., 2013) that might uncover previously unknown physical connectivity, and detailed simulation models (Markram, 2006) that might uncover previously unappreciated biophysical mechanisms for point-like integration of non-local codes.

8.1.2. Consequences from absence of distributed-to-local conversion

If biophysical mechanisms for integration of distributed codes can be definitively excluded by future anatomical and computational investigations, physically orthodox theories cannot explain consciousness. Accordingly, any scientific explanation of consciousness must relax one or more the defining features of physical orthodoxy [5.2]. Most directly, the requirement for locality in physical causation can be dropped. However, this leads to another problem: ensuing non-locality is of a very particular kind, with a high degree of order – certain sets of locations in the brain are connected together computationally, and then connected to specific $\boldsymbol{\rho}$-coordinate locations. Without physical structure to effect these connections, the details of connections appear to be inexplicable, thus rendering consciousness itself beyond scientific explanation.

As noted in [7.2.6], relaxation of the three-spatial-dimensions or conventional-particle-spectrum assumptions (rather than the relaxation of the locality assumption) can lead to theories that both explain the complex order of connections and preserve locality of physical causation. These theories can hypothesize connective physical structures that are not observable via conventional means, and preserve the possibility of a scientific explanation of consciousness. Thus, if biophysical mechanisms for reduction to local codes are excluded, results here indicate that higher-dimensional or non-orthodox particle spectrum theories-of-consciousness should be prioritized, in the first instance.

8.1.3. "Quantum consciousness" and distributed-to-local conversion

Informal discussions of locality issues elsewhere have resulted in observations such as "any physical process responsible for consciousness would have to be something with an essentially global [non-local] character" (Penrose, 2000), and a related inference that consciousness must therefore explicitly depend on quantum mechanisms. More precise, symbol-based, formal considerations in the present paper provide a number of grounds for challenging the alleged *necessity* for an explicitly quantum theory-of-consciousness based on hypothetical non-locality in brain action on a number of grounds. First, it is possible that biophysical mechanisms exist to convert non-local to local codes (and will be observed) [8.1.1]. Second, if these mechanisms are excluded, at least two classical proposals (involving physical structure either in extra dimensions or comprised of non-orthodox particles) exist [8.1.2] as alternatives to "quantum consciousness". Finally, even if quantum non-locality is involved *e.g.* in the integration of spatially-distributed brain-dynamical information, further mechanisms explaining the precise order of connectivity [8.1.2] amongst $\{{}_{\mathrm{B}}\mathbf{r}_i\}$ locations and between $\{{}_{\mathrm{B}}\mathbf{r}_i\}$ and $\{\boldsymbol{\rho}_j\}$ locations must be proposed: these do not follow merely from the hypothesis of a quantum mechanism, and are signally lacking in current quantum theories-of-consciousness [*e.g.* (Penrose, 1989; Stapp, 1993; Beck and Eccles, 2003)].

8.2. Robustness

The exclusion of distributed codes for consciousness is hard to avoid, precisely because the general approach creates few assumptions whose invalidity might falsify reasoning. Note that an objection such as "neural coding of consciousness might not depend on distributed codes" does not refute results: it is simply another way of stating that biophysical mechanisms must exist to reduce distributed or extended codes to spatiotemporally point-like states.

If psychophysical parallelism is excluded [2.3], the only apparent way to avoid the reduction of distributed codes for a locally-causal theory is to deny that conscious experience exists *anywhere*. This is apparently the route taken by Dennett (Dennett, 1991), although this position is, at best, difficult to understand, because it seems to mean that conscious experience itself does not exist. In the present paper, non-existence is excluded by the presumption that each reader can verify for themselves the meaning of the terminology "exteroceptive, visual and auditory, conscious experience" [2.1,2.2].

8.3. Outlook

Central results here help to shape the outlook for the scientific understanding of human consciousness. Because neither spatially-distributed nor temporally-extended codes can be physically-orthodox informational bases for consciousness, point-like encodings and associated biophysical machinery for their generation must be observed if the currently mainstream physically-orthodox presumption is to be sustained, thus establishing for the first time experimental tests for the physically-orthodox hypothesis. Emphatically, tests here do not arbitrate between alternative "hows" for consciousness (such as phase synchrony, or re-entrant processing); instead, they address the more fundamental issue of physical orthodoxy that has until now been amenable only to philosophical (McGinn, 1989; Dennett, 1991; Chalmers, 1996) or even mathematical (Penrose, 1989) analysis.

Experimental outcomes can provide more than (currently absent) empirical support for the physical orthodoxy presumption, however. Observations supporting physical orthodoxy should inform future theoretical and experimental work, by identifying brain locations and physical properties critically involved in the generation of consciousness. Of course, observations rejecting physical orthodoxy would advance the field by ruling out a large class of theories, but theoretical analyses here *also* identify a small group of basic theories that would then become prominent. Notably for the rational conduct of a science of consciousness, even after a putative rejection of physical orthodoxy, the content and multiplicity of theories then remaining rejects allegedly conclusive assertions that consciousness "must" be quantum (Penrose, 1989; Stapp, 1993; Beck and Eccles, 2003), non-physical (Chalmers, 1996), or beyond human understanding (McGinn, 1989), because remaining candidates are all physical (albeit of a non-orthodox kind) and contain no definitively quantum features. Hence, in addition to the characterization of novel and important experiments, developments here can contribute a new dialectical order to the field as whole, helping to instil a clarity that can only be positive for the science of consciousness.

Claims that present methods can provide basic clarity are founded in part on precision due to new formal symbolism for describing phenomena and expressing theoretical alternatives. For example, symbolism has been employed in this paper to treat in precise terms: the delineation between binding problems in behavioral computations (described in A and $\mathbf{r}$ notation) and in conscious experience (B, $\boldsymbol{\rho}$ notation); the relationship between orthodox and conscious-experiential spaces (labelled by $\mathbf{r}$ and $\boldsymbol{\rho}$ symbols respectively); and, both spatial and temporal locality. [(Rosseinsky, 2014b, 2014a) apply symbolism developed here to give a fuller treatment of binding problems.] Notably, a formal approach resolves previously imprecise verbal appeals to emergent properties of complex systems, by revealing the role that intermediate physical structure must play if these appeals are to explain spatially- and temporally-extended phenomena. Precision of this kind is essential if consciousness is to be explained scientifically.

Relatedly, the present approach supports methodological clarity by emphasizing: explicit definitions of phenomenological scope; definitions of physical orthodoxy and its requirements; explicit statement of theories considered and excluded; and explicit demarcation between (large classes of) theories whose comparative status is to be assessed by experiments. Proper treatment of these methodological basics assuages concerns that the study of consciousness is not scientific (Wilkes, 1984, 1988; Hawking, 2000). (Related concerns stemming from the role of subjective report are minimized here by experiments that are entirely *objective* in nature.)

The primary question here has been the conditions under which parallel and distributed codes can generate components-of-consciousness in a manner obeying the local causality requirement of physical orthodoxy. Central results exclude a physically-orthodox encoding-of-consciousness role for spatially-distributed and temporally-extended codes under *any* conditions, but *unconditionally* permit such a role for parallel codes. However, physical orthodoxy places more requirements on theories than just local causality. Subsequent papers in the present series (Rosseinsky, 2014b, 2014a) employ the basic approach developed here to investigate further restrictions on parallel codes imposed by physical orthodoxy, extending the range of experimental tests for physically-orthodox theories-of-consciousness, and providing further demonstration of the positive contributions available from the approach initiated in this paper.

Table 1. Summary of formal symbols.

SYMBOL	DEFINITION	SECTION
s_{ab}	*b*-th instance of *a*-th feature-type (stimulus) in the external environment.	3.1
${}_{X}\mathbf{r}_{j}$	*j*-th location in the external environment for sampling of sensory information	3.3
$<s_{ab}>$	Contribution to conscious experience generated by s_{ab}	5.3
$\boldsymbol{\rho}_{j}$	*j*-th location in conscious experience; information sampled at ${}_{X}\mathbf{r}_{j}$ leads to a contribution-to-consciousness at $\boldsymbol{\rho}_{j}$	5.4
A	Measure of brain activity for behavioral encoding	4.2
B	Measure of brain activity for encoding and final brain-dynamical cause of consciousness	5.5.3
${}_{A}\mathbf{r}_{i}$	*i*-th brain location relevant to *A* measurement	4.4
${}_{B}\mathbf{r}_{i}$	*i*-th brain location relevant to *B* measurement	5.5.3
C_{abj}	Classifier function on *A*-states for behavioral encoding	4.5
D_{abj}	Classifier function on *B*-states for encoding and final brain-dynamical cause of consciousness	5.6
$\exists_{P}$	Denotes existence of a physical object, property, or phenomenon	4.5
$\nexists_{P}$	Denotes absence of a physical object, property, or phenomenon	4.5
$\{\ \dots\}$	Set or collection of …	3.2

9. References

Abbott, L., and Sejnowski, T. J. (1999). *Neural codes and distributed representations: foundations of neural computation*. Cambridge, MA: MIT Press.

Abeles, M. (2011). Cell assemblies. *Scholarpedia* 6, 1505. doi:10.4249/scholarpedia.1505.

Adrian, E. D. (1928). *The basis of sensation: the action of the sense organs*. New York: W. W. Norton & Co.

Alivisatos, A. P., Chun, M., Church, G. M., Deisseroth, K., Donoghue, J. P., Greenspan, R. J., McEuen, P. L., Roukes, M., Sejnowski, T. J., and Weiss, P. S. (2013). The brain activity map. *Science* 339, 1284–1285.

Anderson, P. W. (1972). More is different. *Science* 177, 393–396.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. J., and Edelman, D. B. (2012). Consciousness, biology and quantum hypotheses. *Phys. Life Rev.* 9, 285–294.

Beck, F., and Eccles, J. C. (2003). Quantum processes in the brain: A scientific basis of consciousness. *Adv. Conscious. Res.* 49, 141–166.

Bell, A. J., and Sejnowski, T. J. (1996). The“ independent components” of natural scenes are edge filters. *Vision Res.* 37, 3327–3338.

Buonomano, D. V., and Merzenich, M. M. (1995). Temporal information transformed into a spatial code by a neural network with realistic properties. *Science*, 1028–1030.

Buzsáki, G. (2006). *Rhythms of the brain*. Oxford; New York: Oxford University Press.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chalmers, D. J. (2008). “The varieties of emergence,” in *The Re-emergence of emergence: The Emergentist Hypothesis from Science to Religion*, eds. P. Davies and P. Clayton (Oxford: Oxford University Press), 244–256.

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *J. Philos.* 82, 8–28.

Churchland, P. S. (2005). A neurophilosophical slant on consciousness research. *Prog. Brain Res.* 149, 285.

Churchland, P. S., and Sejnowski, T. J. (1992). *The Computational Brain*. Cambridge, MA: MIT press.

Crick, F., and Koch, C. (1990). Towards a neurobiological theory of consciousness. *Semin. Neurosci.*

2, 263–275.

Decharms, R. C., and Zador, A. (2000). Neural representation and the cortical code. *Annu. Rev. Neurosci.* 23, 613–647.

Dehaene, S. (2014). *Consciousness and the brain: deciphering how the brain codes our thoughts*. New York: Viking.

Dennett, D. C. (1991). *Consciousness Explained*. Boston, MA: Little Brown.

Dennett, D. C., and Kinsbourne, M. (1992). Time and the Observer: the Where and When of Consciousness in the Brain. *Behav. Brain Sci.* 15, 183–247.

Edelman, G. M. (1989). *The remembered present: A biological theory of consciousness*. New York: Basic Books.

Eggermont, J. J. (1998). Is there a neural code? *Neurosci. Biobehav. Rev.* 22, 355–370.

Engel, A. K., and Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends Cogn. Sci.* 5, 16–25.

Faisal, A. A., Selen, L. P. J., and Wolpert, D. M. (2008). Noise in the nervous system. *Nat. Rev. Neurosci.* 9, 292–303.

Feldman, J. (2013). The neural binding problem(s). *Cogn. Neurodyn.* 7, 1–11.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Comput.* 6, 559–601.

Foley, J. M. (1978). “Primary distance perception,” in *Perception* (Berlin: Springer), 181–213.

Foley, J. M., Ribeiro-Filho, N. P., and Da Silva, J. A. (2004). Visual perception of extent and the geometry of visual space. *Vision Res.* 44, 147–156.

Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science* 233, 1416–1419.

Van Gulick, R. (2001). Reduction, emergence and other recent options on the mind/body problem. A philosophic overview. *J. Conscious. Stud.* 8, 1–34.

Hameroff, S. (2012). How quantum brain biology can rescue conscious free will. *Front. Integr. Neurosci.* 6. doi:10.3389/fnint.2012.00093.

Hawking, S. W. (2000). “The Objections of an Unashamed Reductionist,” in *The Large, the Small and the Human Mind*, ed. M. Longair (Cambridge: Cambridge University Press), 169–172.

Haynes, J.-D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534.

Heller, J. (1997). On the psychophysics of binocular space perception. *J. Math. Psychol.* 41, 29–43.

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural Decoding of Visual Imagery During Sleep. *Science* 340, 639–642. doi:10.1126/science.1234330.

Judd, D. B., and Wyszecki, G. (1975). *Color in business, science and industry*. Chichester: Wiley.

Kaas, J. H. (1997). Topographic maps are fundamental to sensory processing. *Brain Res. Bull.* 44, 107–112.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., and Hudspeth, A. J. eds. (2012). *Principles of Neural Science*. New York: McGraw-Hill Professional.

Klimesch, W., Doppelmayr, M., Yonelinas, A., Kroll, N. E. A., Lazzara, M., Röhm, D., and Gruber, W. (2001). Theta synchronization during episodic retrieval: neural correlates of conscious awareness. *Cogn. Brain Res.* 12, 33–38.

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends Cogn. Sci.* 16, 317–321.

Luneberg, R. K. (1944). Mathematical theory of optics (lecture notes). *Brown Univ.*

Mainen, Z. F., and Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science* 268, 1503–1506.

Von der Malsburg, C. (1981). The Correlation Theory of Brain Function. Max Planck Institute for Biophysical Chemistry, Dept. of Neurobiology.

Markram, H. (2006). The blue brain project. *Nat. Rev. Neurosci.* 7, 153–160.

McFadden, J. (2002). The Conscious Electromagnetic Information (Cemi) Field Theory: The Hard Problem Made Easy? *J. Conscious. Stud.* 9, 45–60.

McGinn, C. (1989). Can We Solve the Mind–Body Problem? *Mind* 98, 349–366.

Mehta, M. R., Lee, A. K., and Wilson, M. A. (2002). Role of experience and oscillations in transforming a rate code into a temporal code. *Nature* 417, 741–746. doi:10.1038/nature00807.

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

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behav. Brain Sci.* 30, 63–81.

Miller, G. (2005). What Is the Biological Basis of Consciousness? *Science* 309, 79. doi:10.1126/science.309.5731.79.

Olshausen, B. A. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609.

Panzeri, S., Brunel, N., Logothetis, N. K., and Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. *Trends Neurosci.* 33, 111.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

Penrose, R. (2000). *The Large, the Small and the Human Mind*. Cambridge: Cambridge University Press.

Perkel, D. H., and Bullock, T. H. (1968). Neural coding. *Neurosci. Res. Program Bull.* 6, 219–349.

Phillips, C. G., Zeki, S., and Barlow, H. B. (1984). Localization of function in the cerebral cortex: past, present and future. *Brain* 107, 328–361.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 454, 995–999.

Place, U. T. (1956). Is consciousness a brain process? *Br. J. Psychol.* 47, 44–50.

Pockett, S. (2002). Difficulties with the electromagnetic field theory of consciousness. *J. Conscious. Stud.* 9, 51–56.

Popper, K. R., and Eccles, J. C. (1977). *The self and its brain*. Berlin: Springer International.

Quiroga, R. Q., and Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nat. Rev. Neurosci.* 10, 173–185.

Reinagel, P., and Reid, R. C. (2000). Temporal Coding of Visual Information in the Thalamus. *J. Neurosci.* 20, 5392–5400.

Rieke, F., de Ruyter van Steveninck, R., Warland, D., and Bialek, W. (1997). *Spikes: exploring the neural code*. Cambridge, MA: MIT Press.

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

Rosseinsky, N. M. (2014a). Multi-subject settings restrict and define physically-orthodox neural codes for consciousness. Manuscript in Submission.

Rosseinsky, N. M. (2014b). Topographic codes for consciousness are constrained to a small and distinctive set by a physically-orthodox setting. Manuscript in Submission.

Rumelhart, D. E., and McClelland, J. L. (1988). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vols. 1 and 2. Cambridge, MA: Bradford.

Sejnowski, T. J., and Paulsen, O. (2006). Network oscillations: emerging computational principles. *J. Neurosci.* 26, 1673–1676.

Shulman, R. G., Hyder, F., and Rothman, D. L. (2009). Baseline brain energy supports the state of consciousness. *Proc. Natl. Acad. Sci.* 106, 11096–11101.

Smart, J. J. C. (1959). Sensations and brain processes. *Philos. Rev.* 68, 141–156.

Solomon, S. G., and Lennie, P. (2007). The machinery of colour vision. *Nat. Rev. Neurosci.* 8, 276–286.

Stapp, H. P. (1993). *Mind, matter, and quantum mechanics*. Berlin: Springer.

Tegmark, M. (2000). Importance of quantum decoherence in brain processes. *Phys. Rev. E* 61, 4194.

Thorpe, S. J., and Imbert, M. (1989). “Biological constraints on connectionist modelling,” in *Connectionism in perspective*, 63–92.

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

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

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

Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. *Neuron* 24, 105–110.

Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J., and McNaughton, B. L. (1996). Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. *Hippocampus* 6, 271–280.

VanRullen, R., and Koch, C. (2003). Is perception discrete or continuous? *Trends Cogn. Sci.* 7, 207–213.

Van Vreeswijk, C. (2004). “What is the neural code,” in *23 Problems in Systems Neuroscience*, eds. J. L. van Hemmen and T. J. Sejnowski (New York: Oxford University Press), 143–159.

Wagner, M. (2006). *The Geometries of Visual Space*. New York: Psychology Press.

Wilkes, K. V. (1988). “---, yishi, , duh, um, and consciousness,” in *Consciousness and Contemporary Science*, eds. A. Marcel and E. Bisiach (Oxford: Clarendon Press), 16–41.

Wilkes, K. V. (1984). Is consciousness important? *Br. J. Philos. Sci.* 35, 223–243.

Zaidi, Q., Victor, J., McDermott, J., Geffen, M., Bensmaia, S., and Cleland, T. A. (2013). Perceptual Spaces: Mathematical Structures to Neural Mechanisms. *J. Neurosci.* 33, 17597–17602. doi:10.1523/jneurosci.3343-13.2013.

10. Figures

Figure 1. Schematic illustration of the primary result. (A) Distributed vector encoding *e.g.* of color *as a component-of-consciousness* leads to non-locality under physical orthodoxy, because a property (conscious experience of color) depends on dynamical activity at more than one location. In a physically-orthodox theory-of-consciousness there can be no intermediate, dedicated, physical structure connecting encoding locations to conscious-experiential locations. (Double-headed black arrow schematically indicates that the red component-of-consciousness is a property of or synonymous with encoding dynamics at multiple locations.) **(B)** Distributed vector encoding *e.g.* of color *as part of conventional computational neuroscience* does not lead to non-locality, because later computational sites are connected to each vector component via intermediate physical structure (*e.g.* axons, synapses, and dendrites). The frontal spike shown, encoding *e.g.* "existence of red in the visual field", is generated by a feedforward computation that is local at each point. (Single-headed black arrow schematically depicts feedforward axon-synapse-dendrite connectivity between encoding dynamics at multiple locations and "behavioral computation" at a single location.)

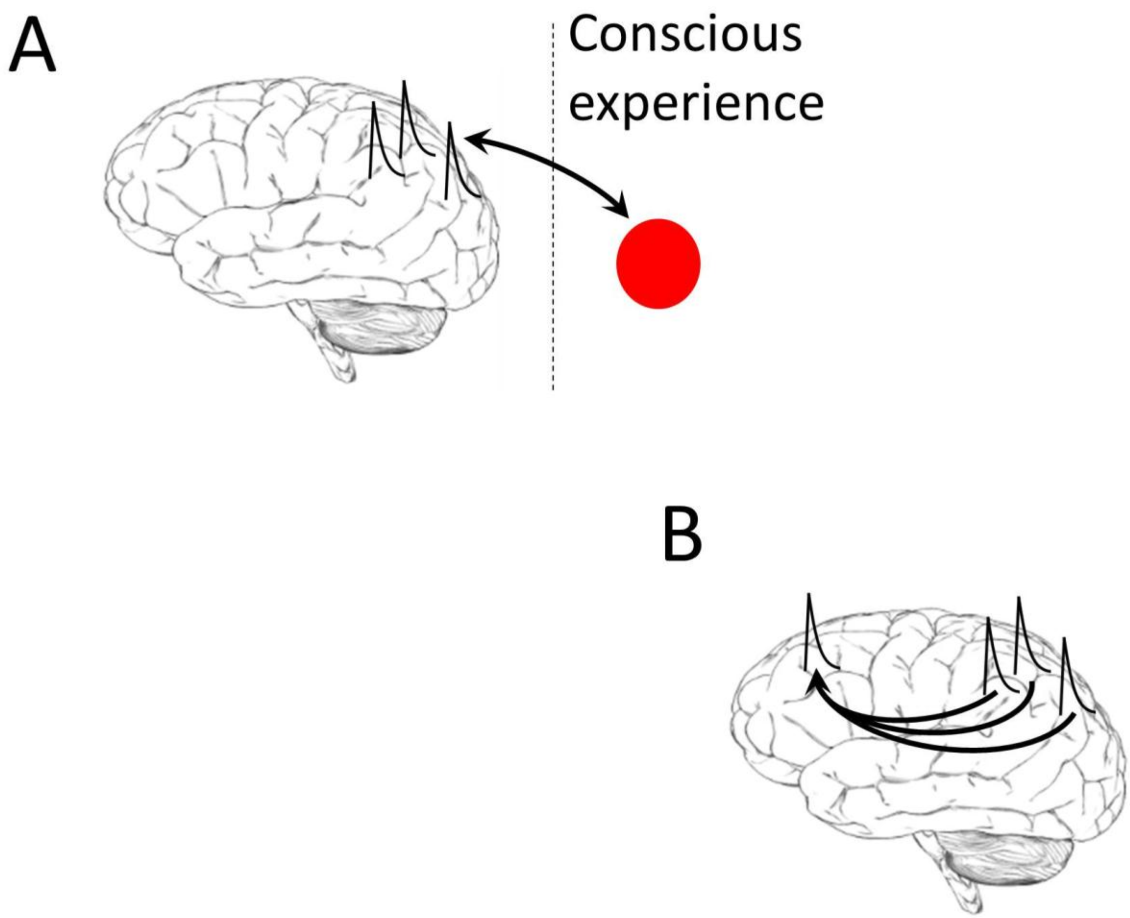

Figure 2. Schematic illustration of "parallel" and "distributed" nomenclature. Three distinct locations in the visual field are schematically identified by colored crosses. Two topographic maps in cortical areas "1" and "2" encode different features of the visual field (oriented edges and motion respectively, say). Colored cortical locations denote neurons in each cortical area dedicated to encoding the correspondingly-colored location in the visual environment. Specifically, *one* neuron per environmental location is used in area 1, and *three* neurons per location in area 2. In this encoding scheme, feature encoding is "parallel" (in two areas), and encodings of distinct locations are computed "in parallel" (in dedicated neurons or groups of neurons, arranged topographically). Area 1 uses "non-distributed" encoding of feature instances, whereas area 2 employs a "distributed" code.

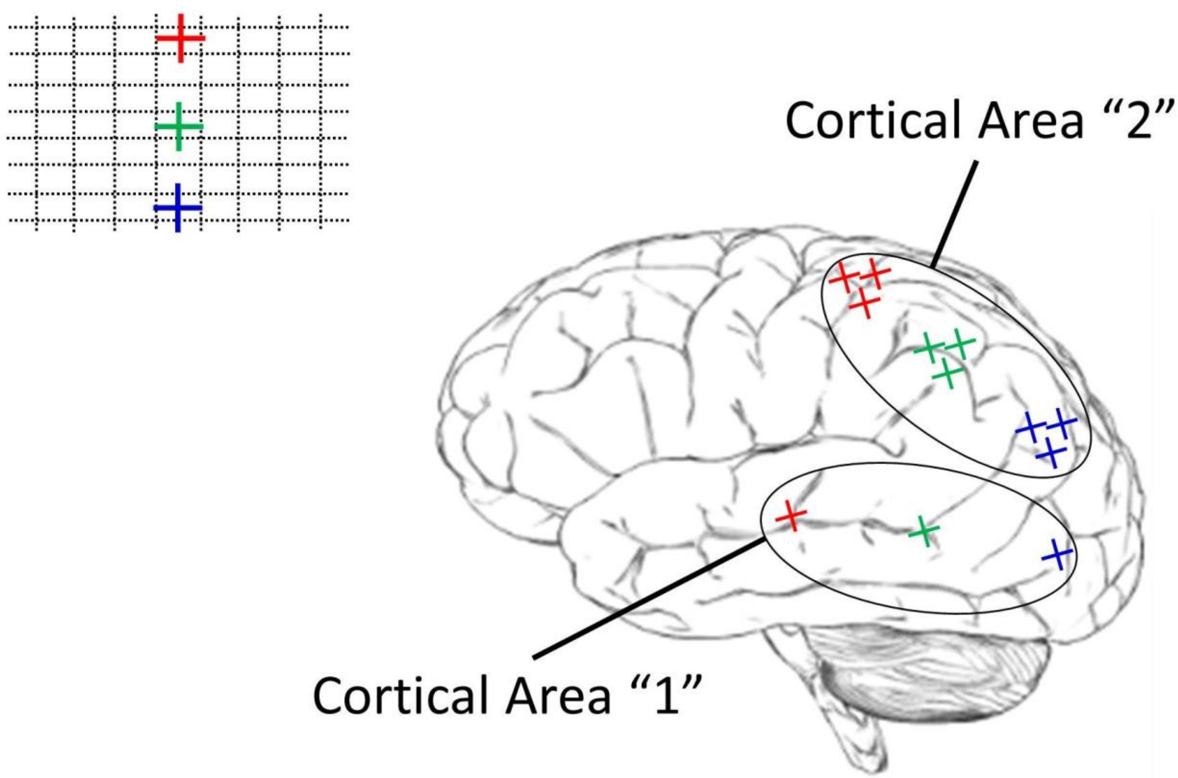

Figure 3. Schematic depiction of locality-of-causality conceptions in multiple-space settings. In this Figure, the light grey space (solid boundary) is identified with the orthodox space containing the physical universe, and the dark grey space (dashed boundary) with the space containing the conscious experience associated with a single brain (not to scale). **(A)** One space embedded in another: two views of the same embedding are depicted, showing that points inside the dashed boundary belong to both the dark grey (upper image) and light grey (lower image) spaces. **(B),(C)** In the embedded construction of panel A, a neural dynamic at Y can either be non-local (panel B) or local (panel C) to a component of conscious experience Z. **(D)** Two topologically-unconnected spaces: conscious-experiential space is not spatially "inside" the physical universe (although physical duality can be avoided by proposing that conscious-experiential *space* is a property of brain dynamics). **(E)** In the unconnected construction of panel D, a neural dynamic at Y cannot be (conventionally) local to a component-of-consciousness at Z, but other neural dynamics U and V can be respectively non-local and local to Y. Thus, two distinct locality conceptions exist, *i.e.* locality of Z to {U,V,Y} and locality of {U,V,Y} to each other.

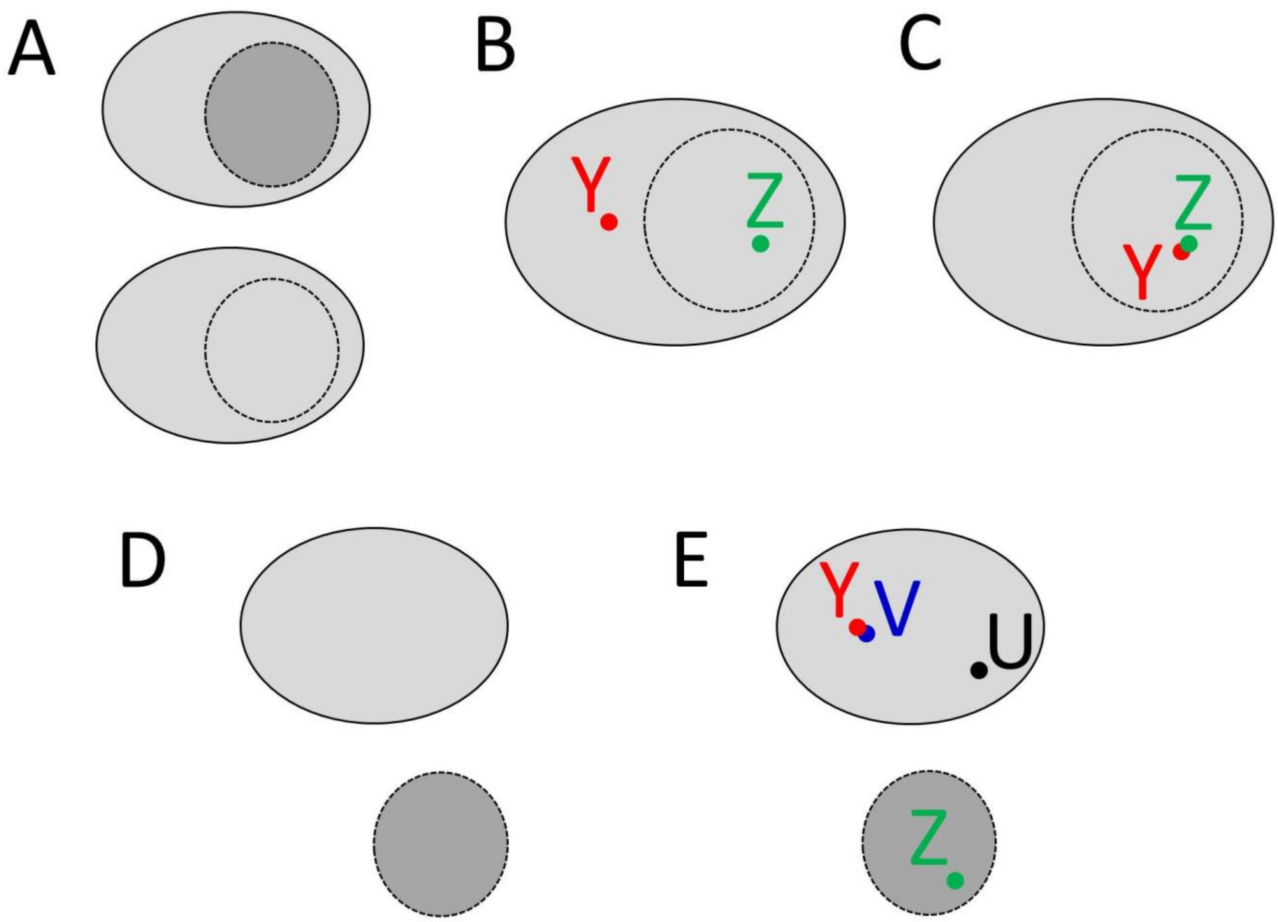

Figure 4. Schematic depiction of phenomena. Central analyses discuss the relationship between the exteroceptive environment (schematically depicted by the line drawing of the tree), brain dynamics that encode the environment (schematically depicted by spikes), and the conscious experience of the environment (schematically depicted by the colored tree and surroundings). The dotted line delineates objective and subjective phenomena without asserting duality: a wide variety of attributions of components-of-conscious-experience to objective brain states will be accommodated by formal symbolism. The central focus will be on *physically-orthodox* attributions that avoid duality while maintaining the necessary recognition of three basic categories of phenomena: environment; brain-dynamical states encoding environment; and, conscious-experience-of-environment.

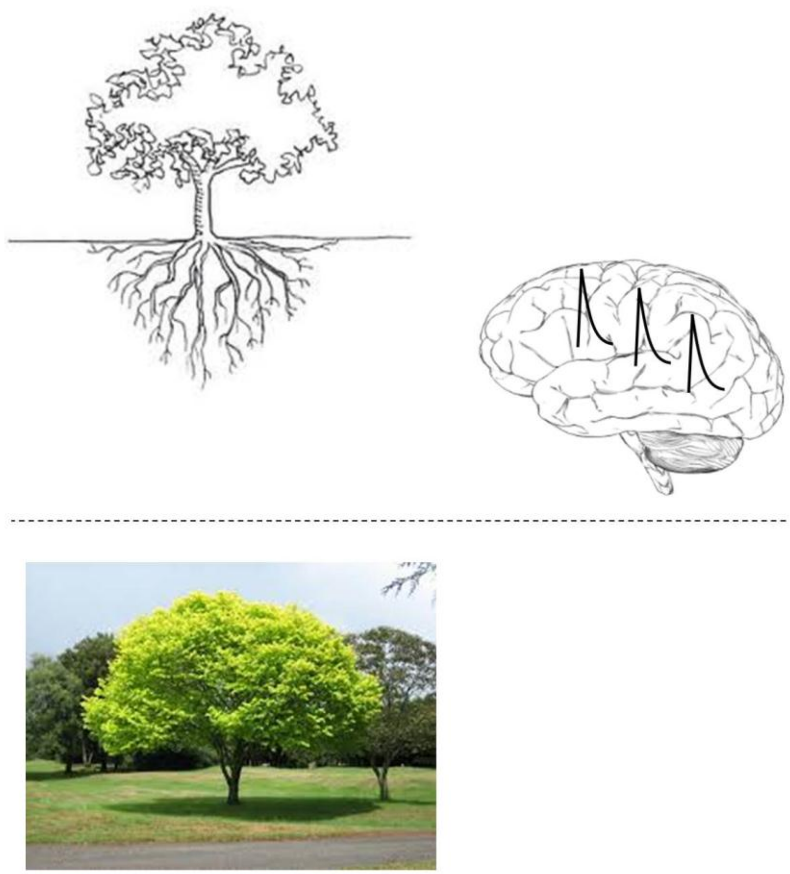

Figure 5. Symbols for components of the exteroceptive environment. Symbols $s_{a,b}$ correspond to specific features of the environment that are primitives (*i.e.* irreducible informational components) of the sensory information-processing systems. For example, each oriented edge shown here corresponds to a specific symbol, $s_{edge,i}$ say. Other members (not shown) of the complete collection $\{s_{a,b}\}$ for the visual modality might include $s_{col,j}$ (color) and $s_{move,k}$ (motion). (Blue double-headed arrows in Figures 5 to 9 schematically depict relationships between natural phenomena and symbols that must be established as part of any scientific investigation.)

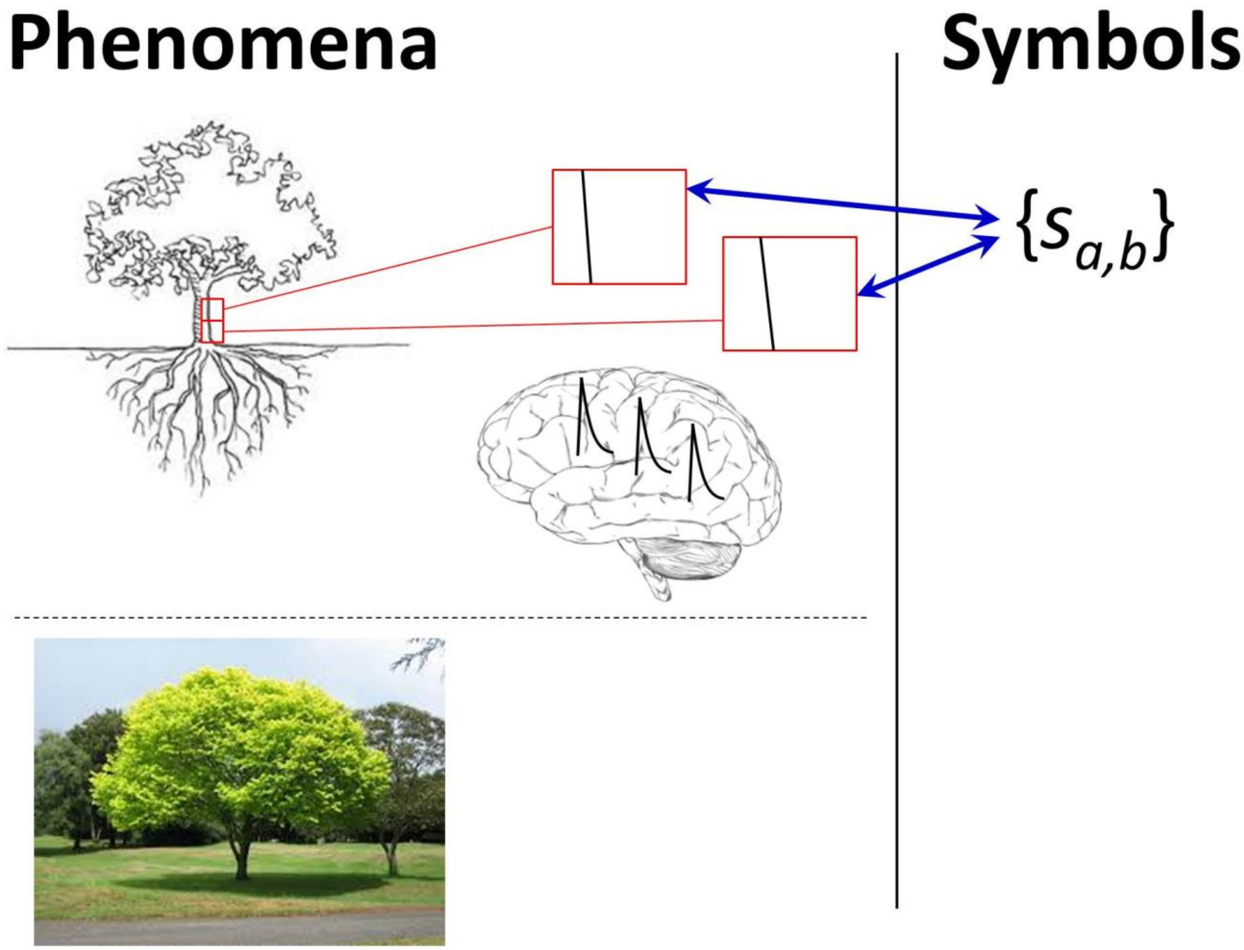

Figure 6. Coordinate symbols for locations in the exteroceptive environment. Symbols ${}_{X}\mathbf{r}_j$ label locations in the exteroceptive environment at which various features labeled by $s_{a,b}$ symbols (Figure 5) occur.

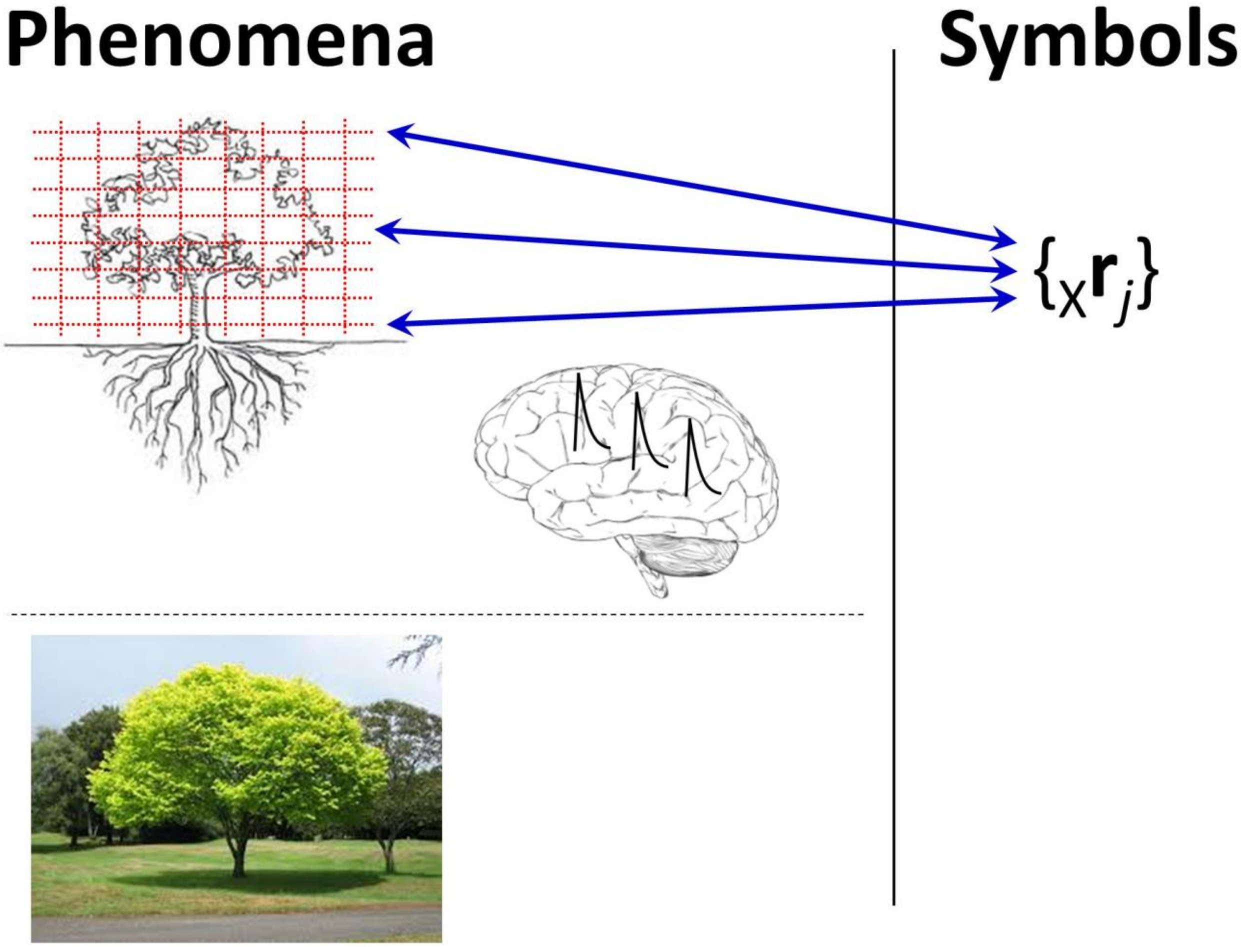

Figure 7. Symbols for brain-dynamics and coordinates for brain locations. The symbol *A* denotes an encoding-relevant measure of brain-dynamical activity. Knowledge of *A* at the locations (red crosses) indexed by the collection of coordinates $\{{}_A\mathbf{r}_i\}$ is sufficient to define the complete encoding state of the brain (for behavioral computations, but not necessarily for the generation of consciousness).

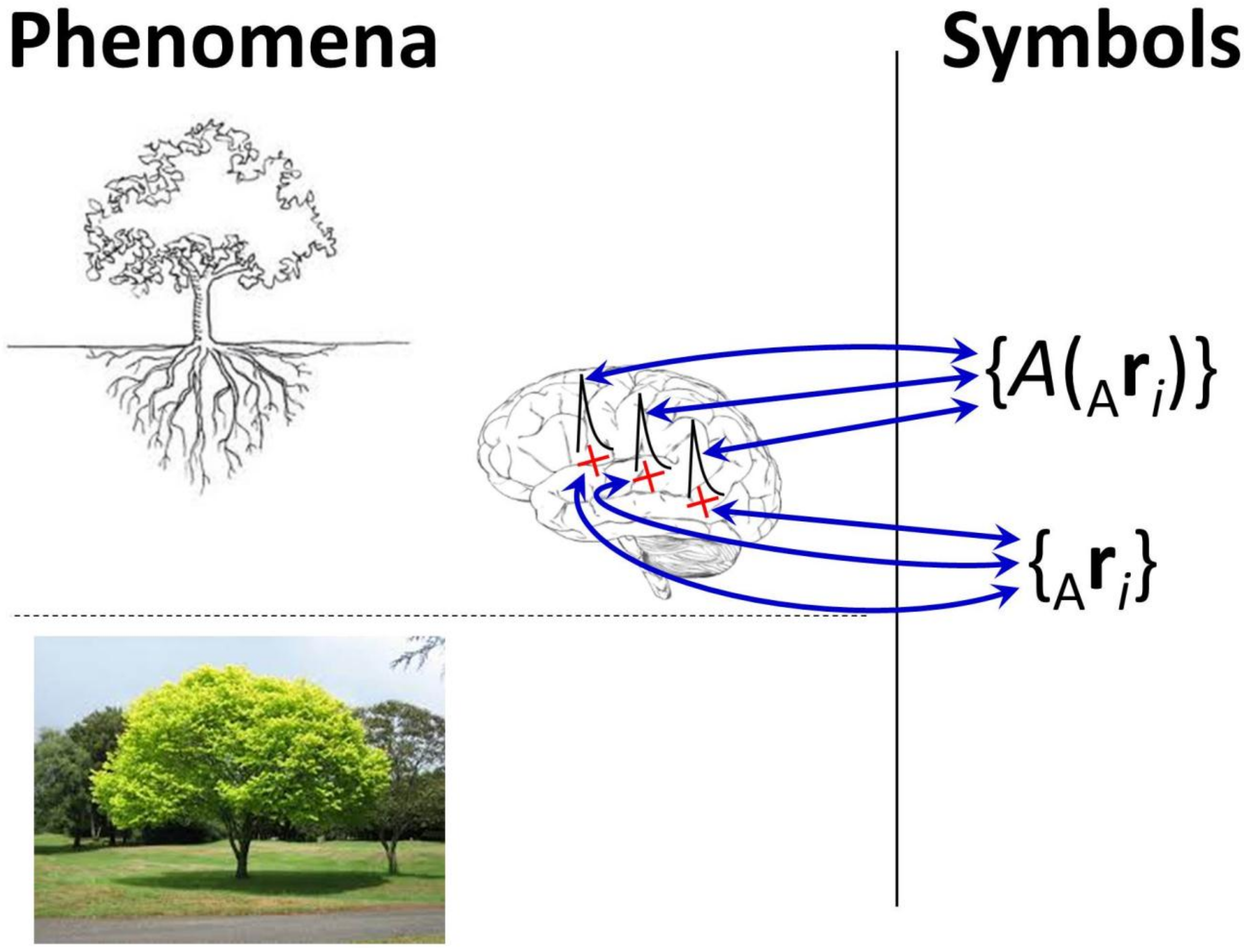

Figure 8. Symbols for components of conscious experience. The symbol <$s_{a,b}$> denotes the conscious experience of the exteroceptive feature labeled by $s_{a,b}$ (Figure 5). (Components of conscious experience shown here are composite combinations of edges and colors. In an $\{s_{edge}, s_{col}\}$ description of the environment, conscious experience would be described in terms of collections of <s_{edge}> and <s_{col}> symbols. Curly parentheses {…} always denote a set or collection, in the present series.)

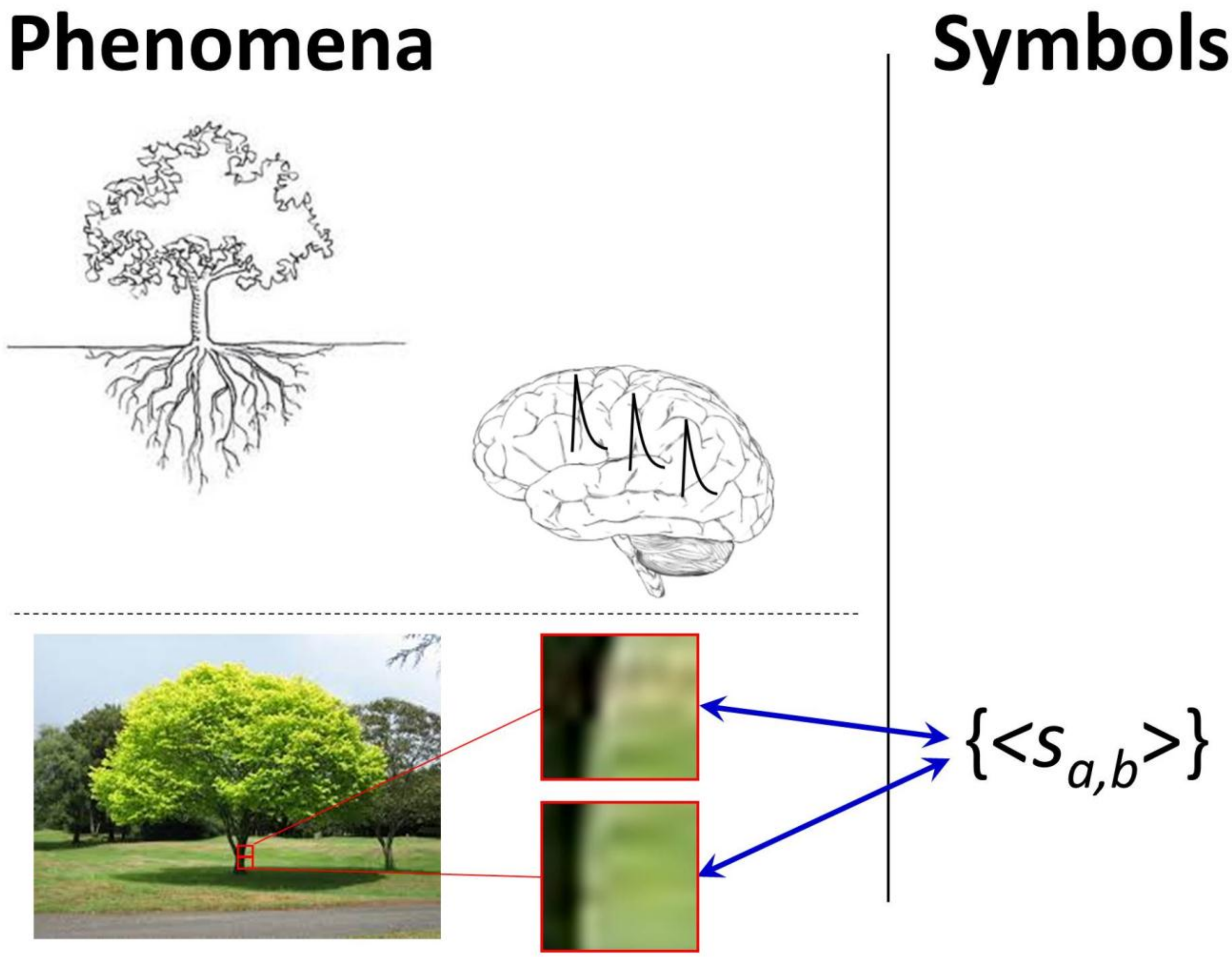

Figure 9. Coordinate symbols for locations in conscious experience. Symbols ρ_k label relative locations within conscious-experiential space at which various components-of-consciousness labeled by $<s_{a,b}>$ symbols (Figure 8) occur.

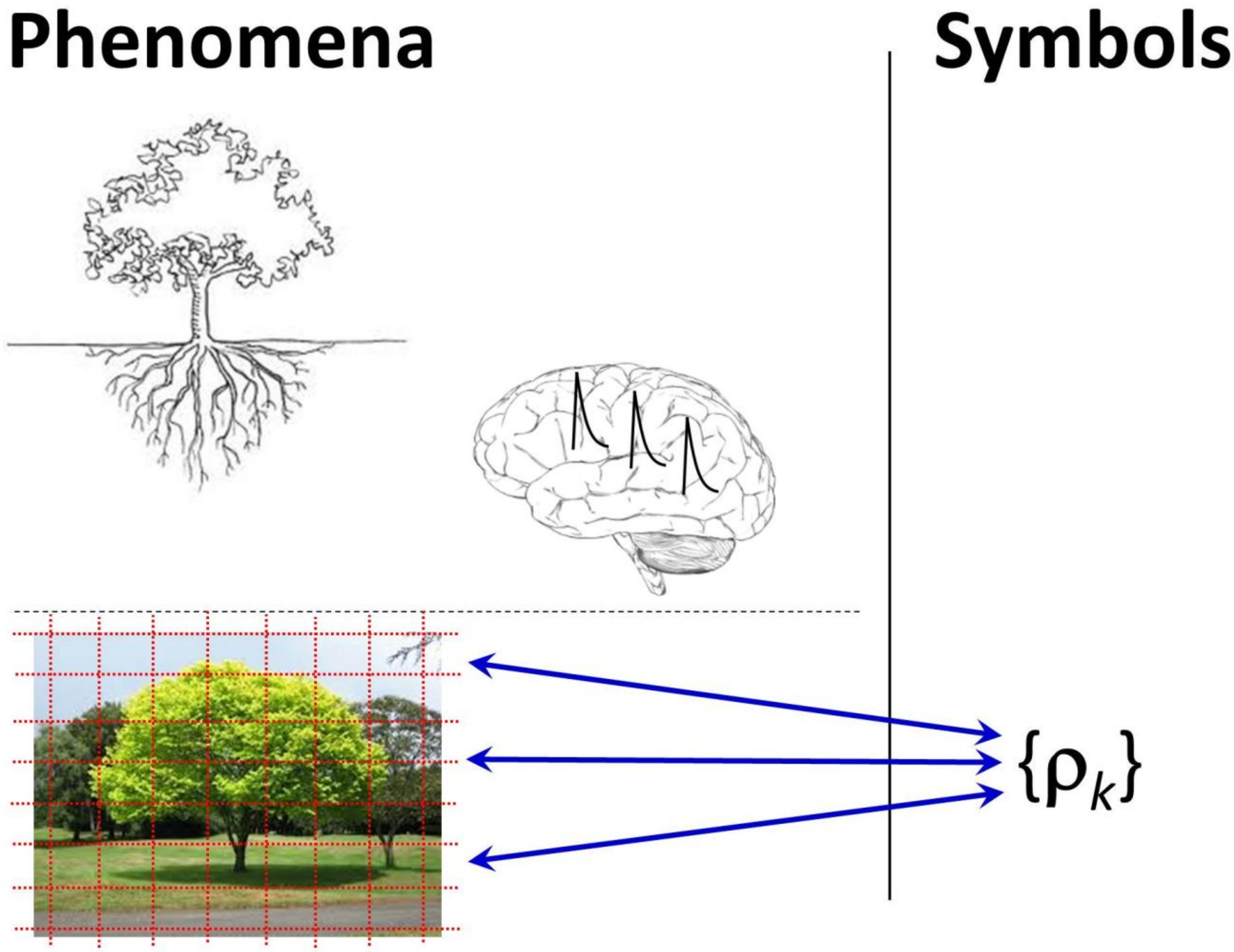

Figure 10. Schematic illustration of the formal statement of a complete theory of exteroceptive consciousness. Extensive definitions of symbols (Figures 4 to 9) lead to just three basic equations that can together completely state any theory-of-consciousness. (Encoding *B*-dynamics are schematically represented by black crosses. Arrows schematically depict causal relationships.) Red arrow, red equation: the existence of a stimulus instance s_{ab} at an external location $_{X}\mathbf{r}_j$ means that *A*-dynamics at brain locations $\{_{A}\mathbf{r}_i\}$ must satisfy $C_{abj} = 1$. Green line, green equation: physical coupling of *A*- and *B*-dynamics means that $C_{abj} = 1$ satisfaction generates *B*-dynamics that satisfy $D_{abj} = 1$. Black line, black equation: by definition of the causal properties of *B*-states, $D_{abj} = 1$ satisfaction causes the generation of the component-of-consciousness $<s_{ab}>$ at the conscious-experiential location ρ_j.

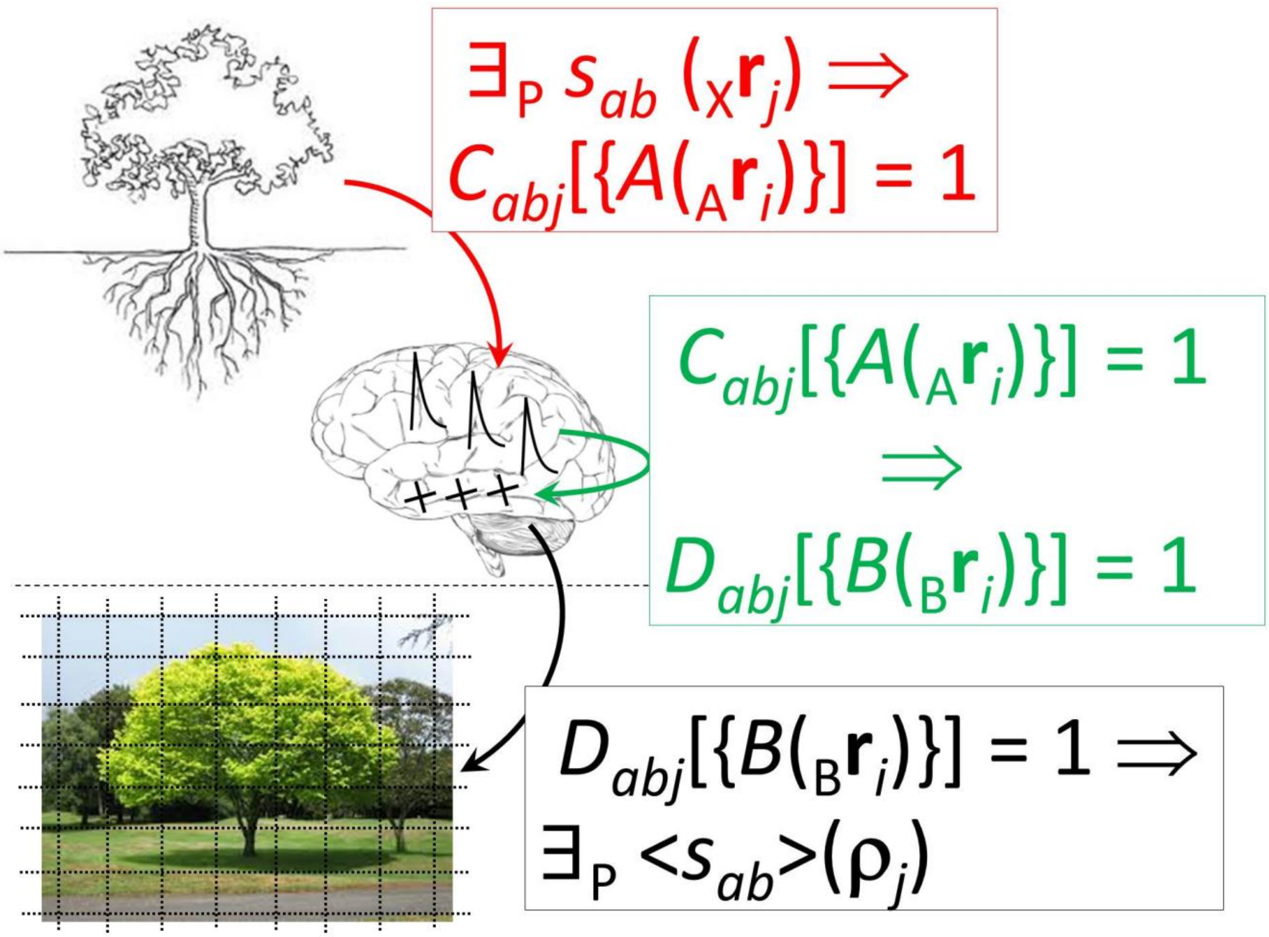

Figure 11. Sub-domain and property schemes for explaining conscious-experiential space. (A),(B) If a sub-domain of orthodox physical space is used as the space of conscious experience (as illustrated in Figure 3A-C), then there is only one space, which can be viewed in two ways: in panel A, neural dynamics encoding experience are displayed as the contents of space; in panel B, experience encoded by dynamics is displayed. **(C)** If both the contents of experience and the space containing that experience are properties of neural dynamics, then the space of conscious experience can be physically and topologically separate from orthodox space (as illustrated in Figure 3D-E). In this case, there are two spaces, one containing brain dynamics (as well as the brain, the environment, and the rest of the physical Universe) and another containing conscious experience. The physically-orthodox avoidance-of-duality can be achieved in this case by considering the space of conscious experience to be ontologically junior to orthodox physical space. Whereas panels A and B depict two different views of *one* space, in the property approach shown in panel C there are two distinct spaces (schematically shown as upper ellipse and lower circle).

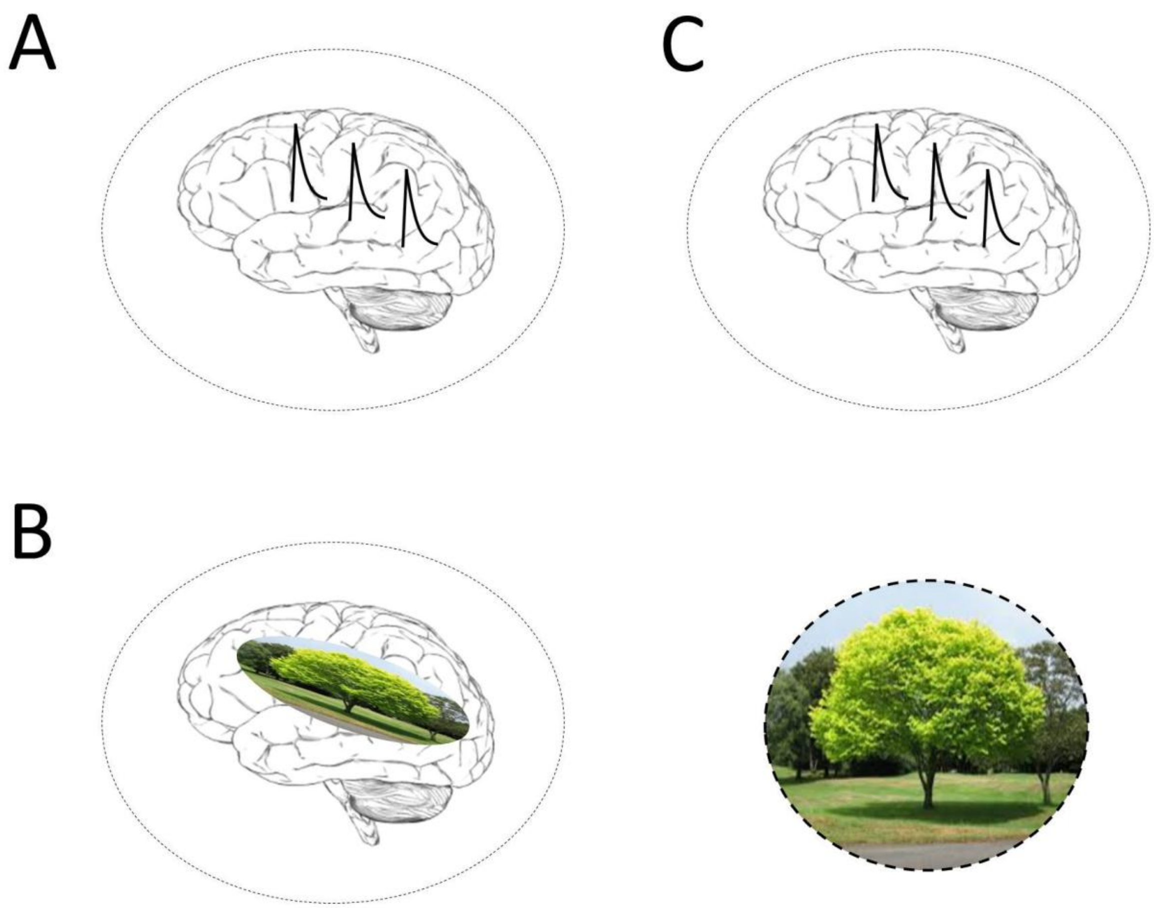

Figure 12. Spatially-extended properties in a local physical theory require a spatially-extended measuring apparatus or other intermediate physical structure. Green picture elements schematically depict physical structure connecting spatially-distributed physical features with spatially-extended properties. Blue double-headed arrows depict connections between physical phenomena and theoretical symbols. Red elements delineate the specific phenomenon labeled by explicitly-displayed symbols. **(A)** A collection of molecules with various momenta (illustrated by arrows of varying direction and magnitude). **(B)** Assignment of temperature T_1 to the collection of molecules requires a spatially-extended thermometer (green; not to scale) whose reading (red line) is the (single-location) phenomenon corresponding to T_1. **(C)** A collection of spatially-distributed neural spikes that together generate a component-of-consciousness (experience of oriented-edge; contents of the red square). **(D)** The label $<s_{12}>(\rho_3)$ assigned to the component-of-consciousness can only be part of a *local* physical theory if there is intermediate physically connective structure (green lines) between neural locations and the location of the component-of-consciousness. This connective structure plays the same conceptual role as the thermometer (panel B) in creating a local theory of a spatially-extended property. Connective structure [7.2.6] could be comprised of *e.g.* exotic particles in a three-dimensional setting, or orthodox particles in a minimally six-dimensional topologically connected setting (Rosseinsky, 2014b). (Other, less orthodox, possibilities for physical connection can be hypothesized.)

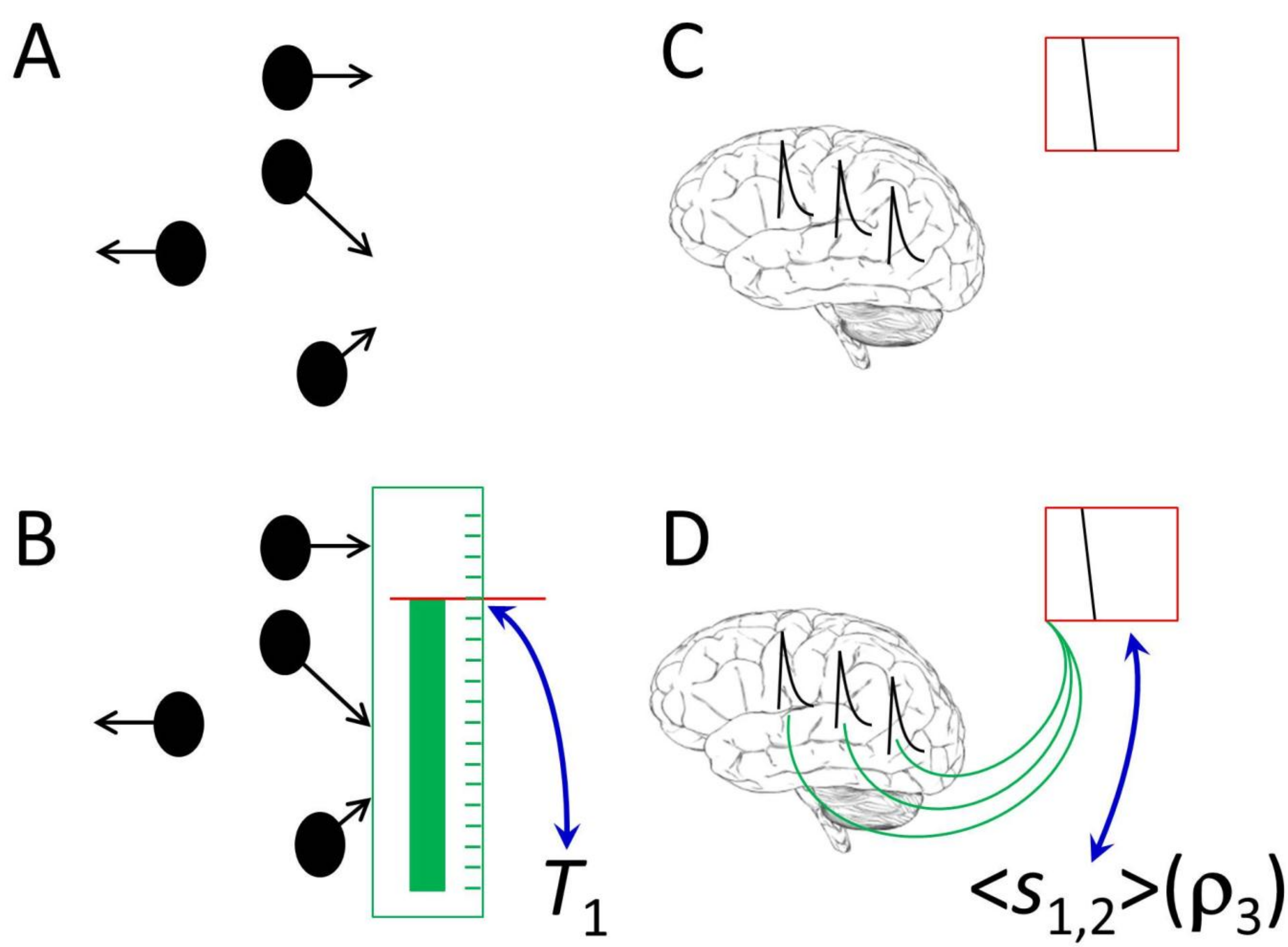

Figure 1.JPEG

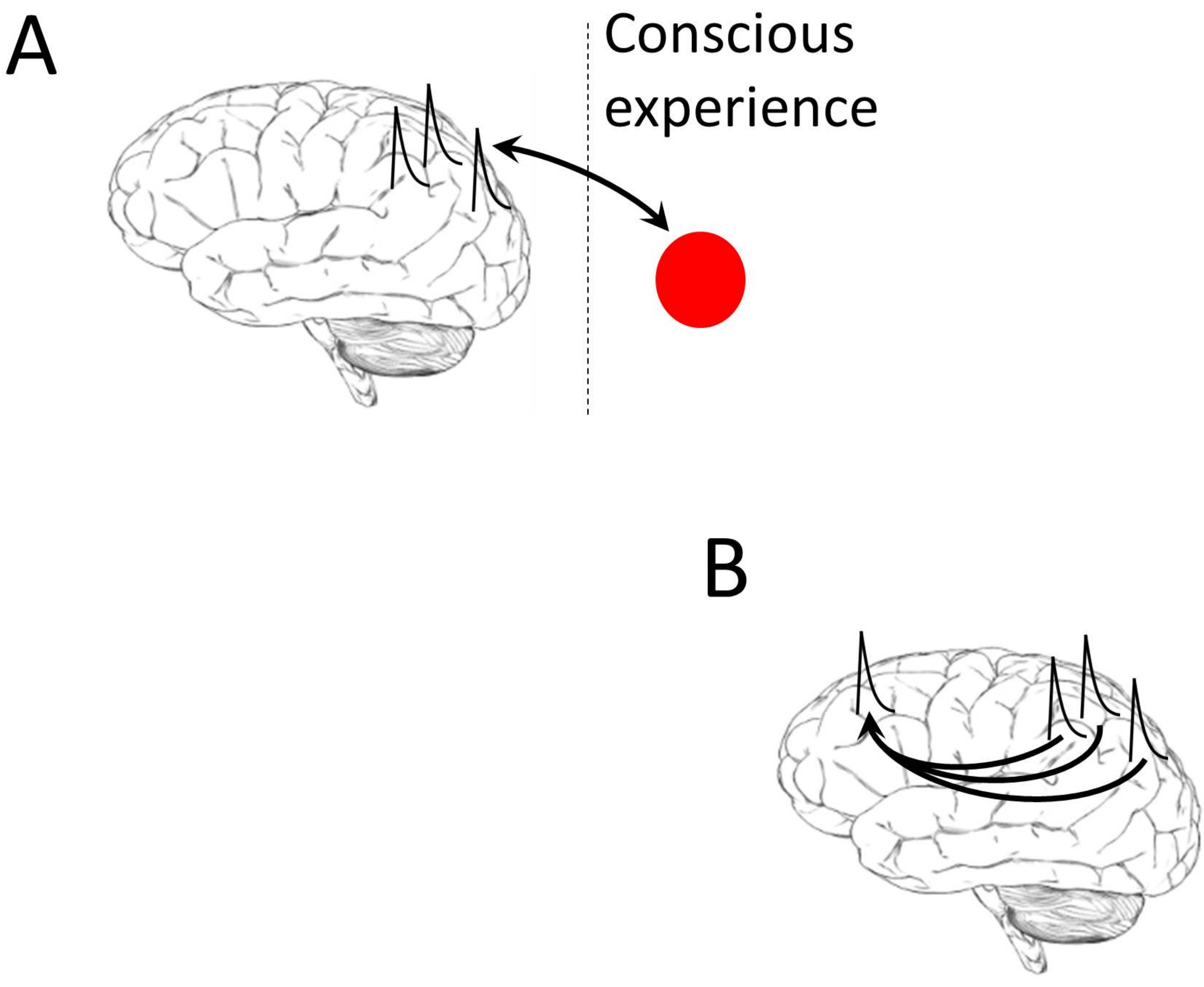

Figure 2.JPEG

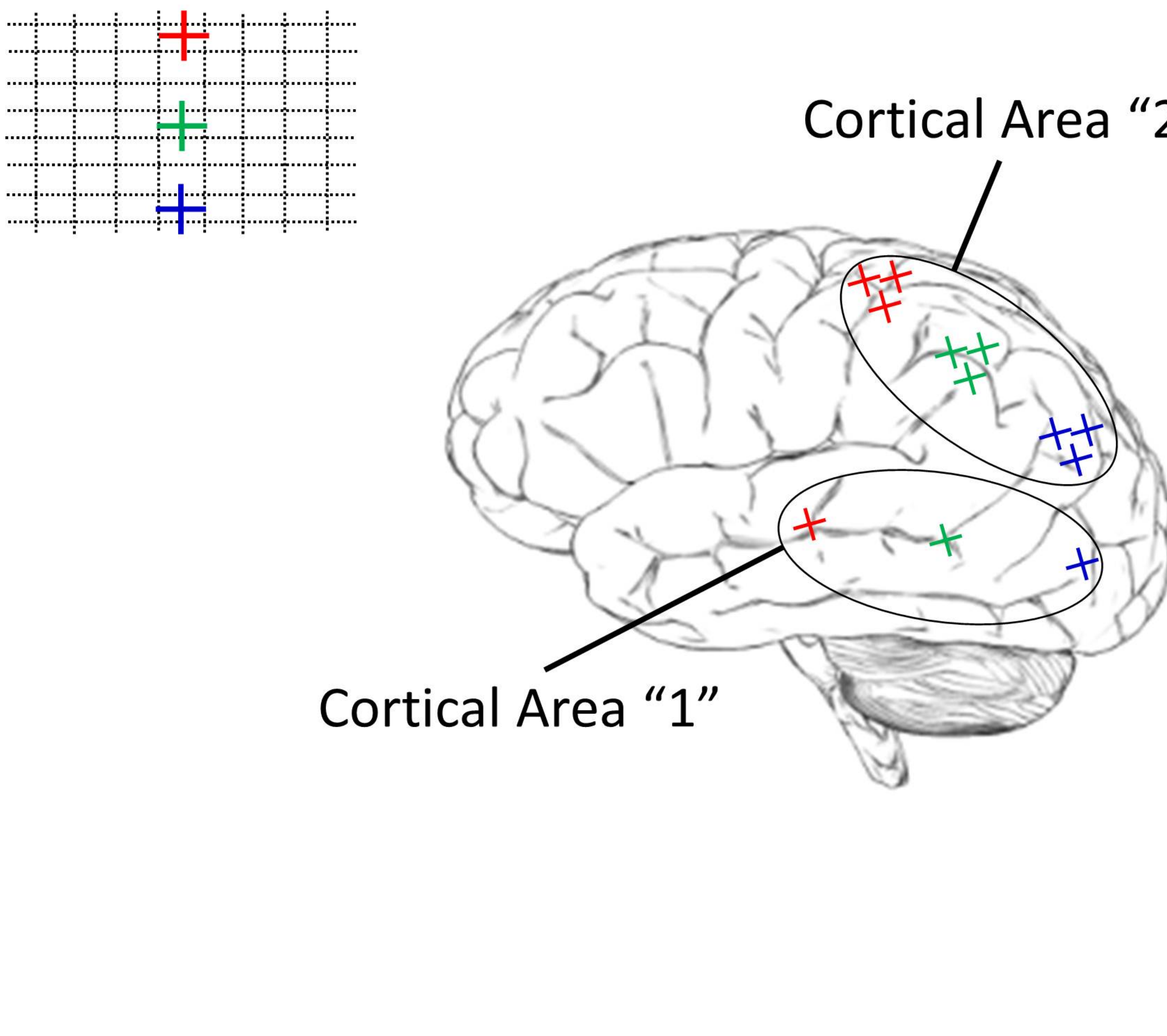

Figure 3.JPEG

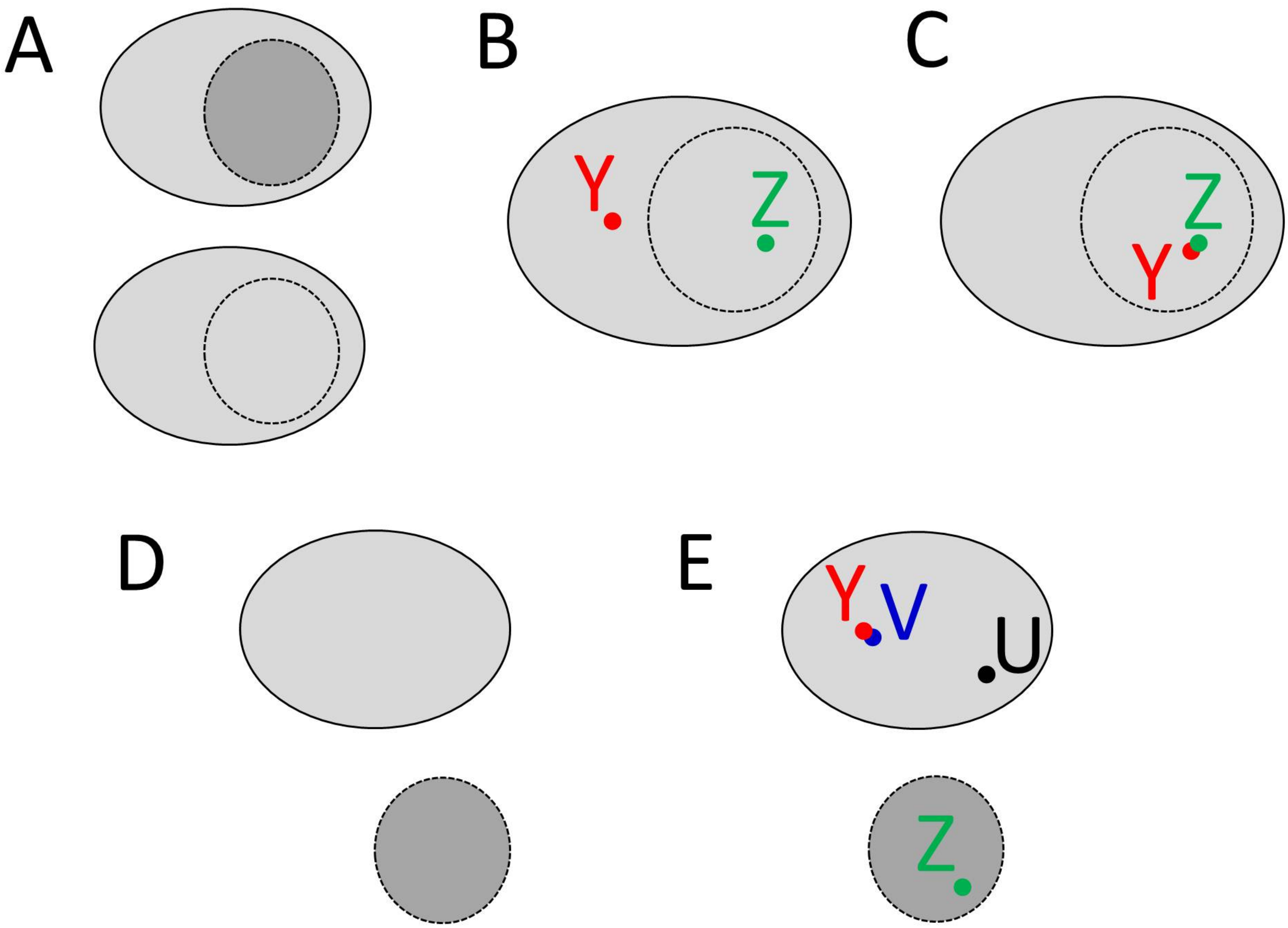

Figure 4.JPEG

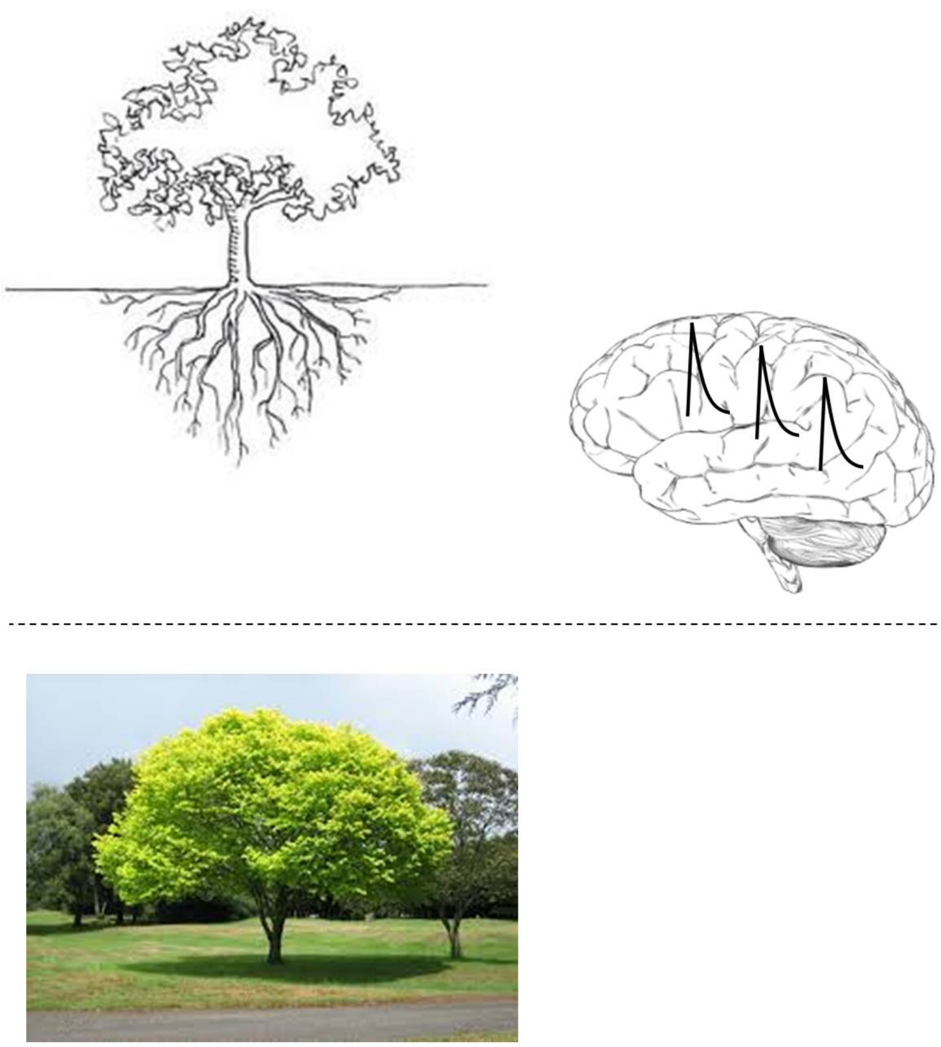

Figure 5.JPEG

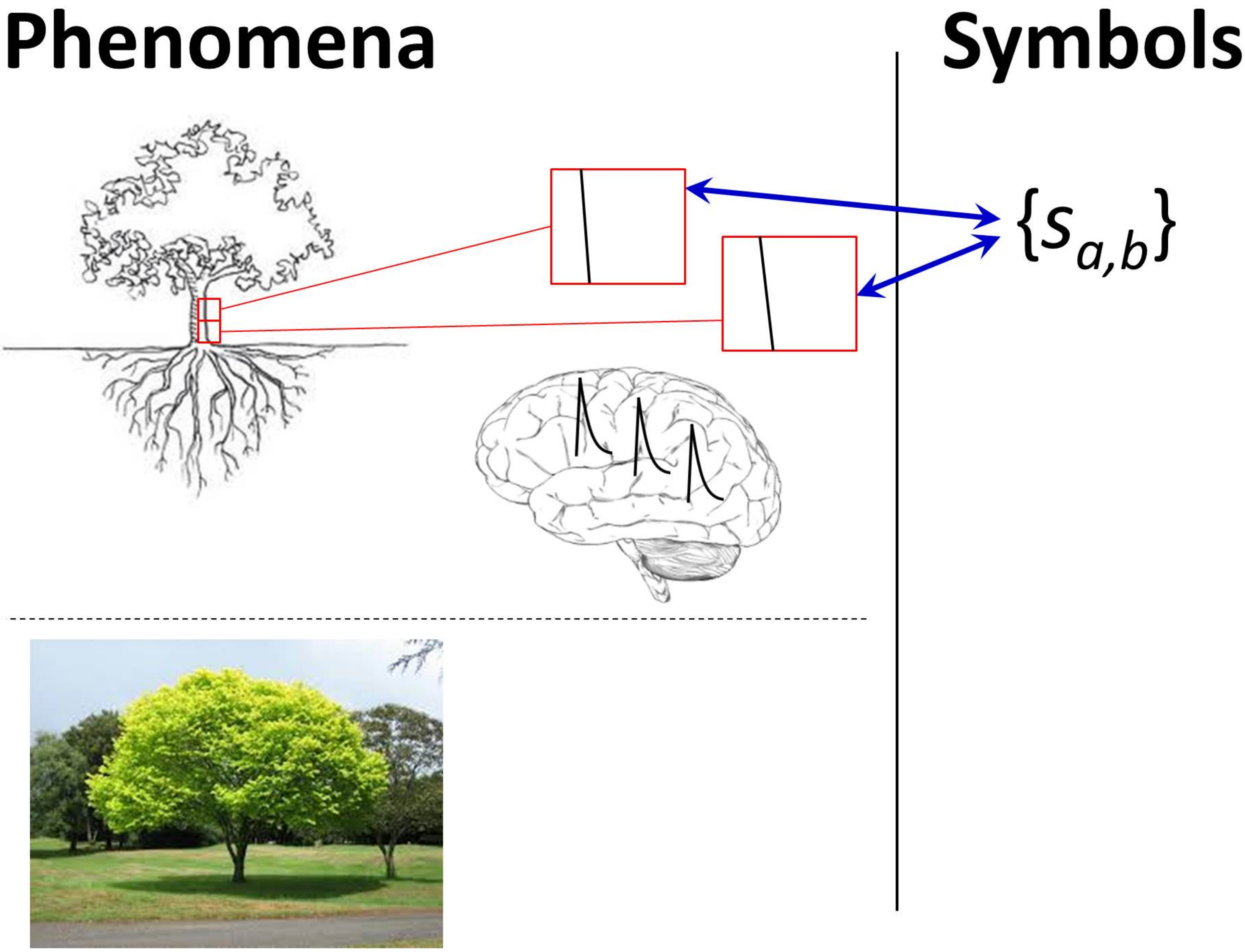

Figure 6.JPEG

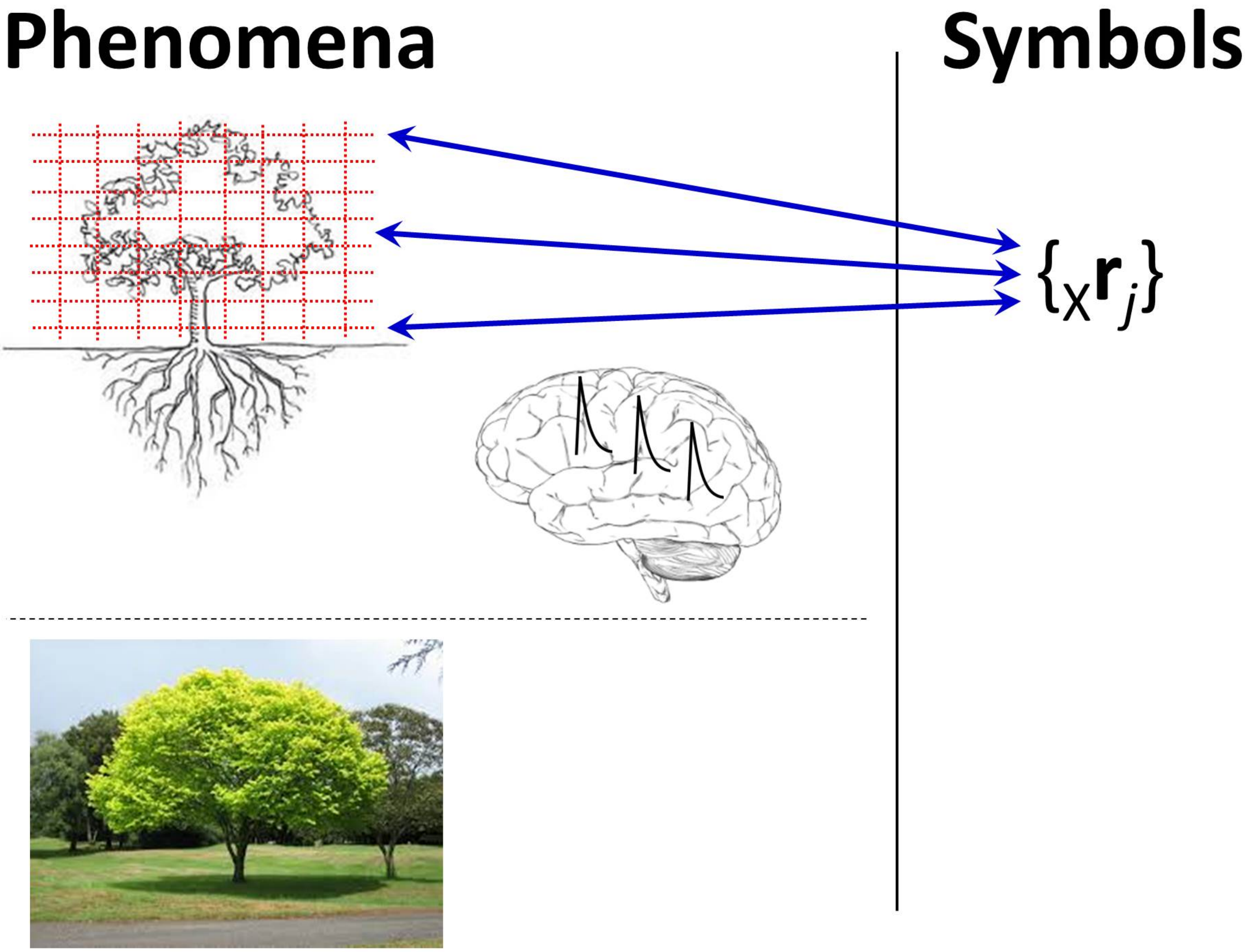

Figure 7.JPEG

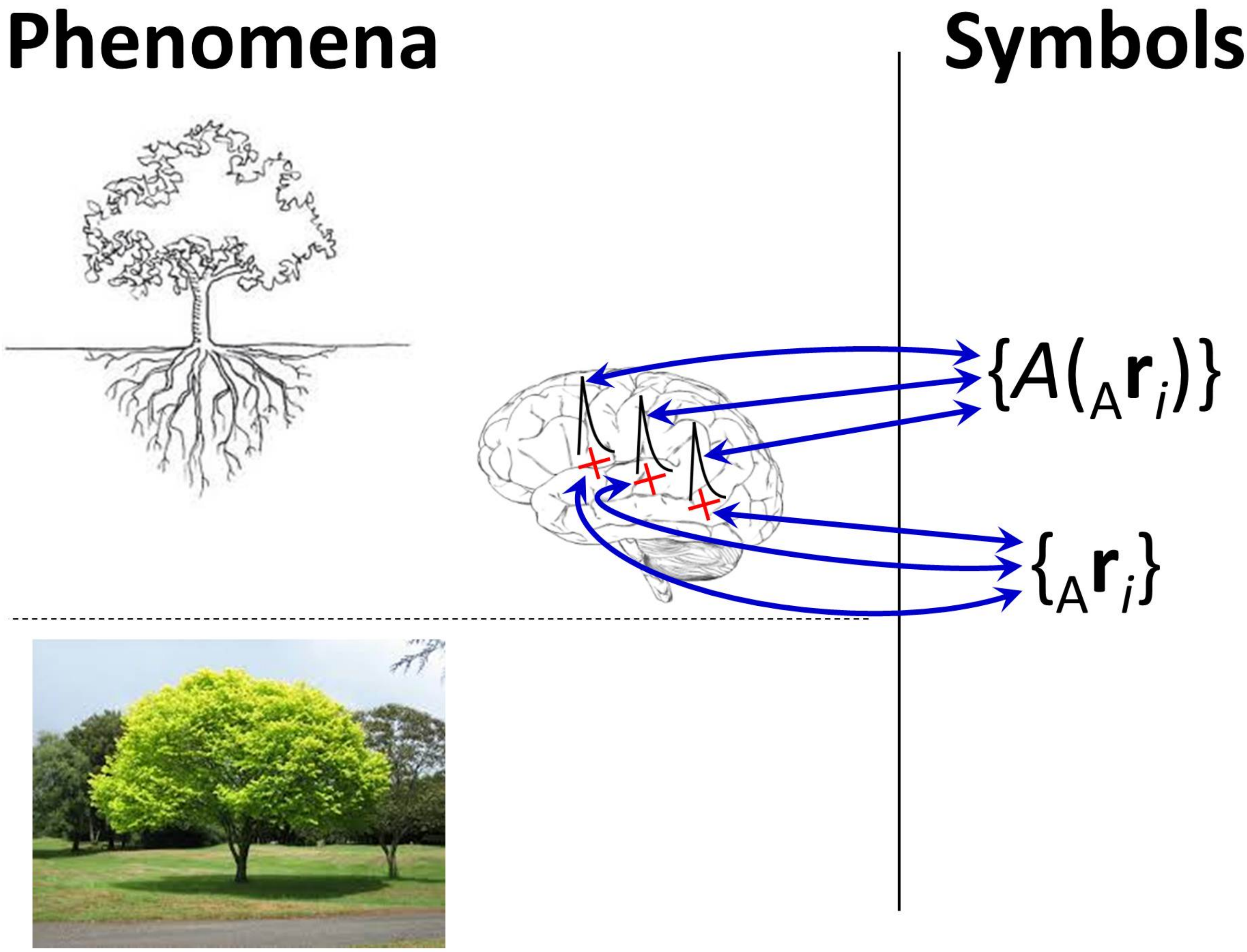

Figure 8.JPEG

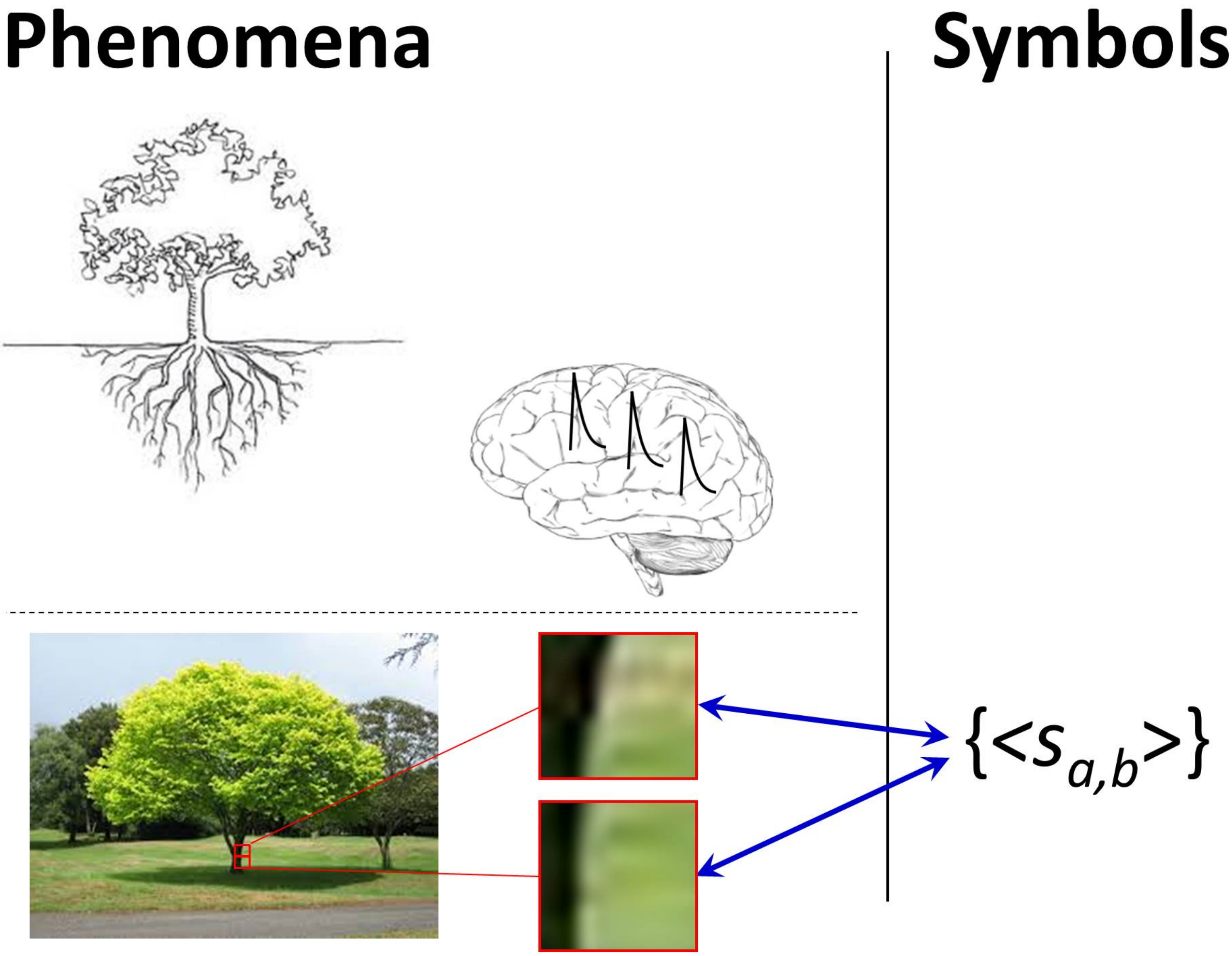

Figure 9.JPEG

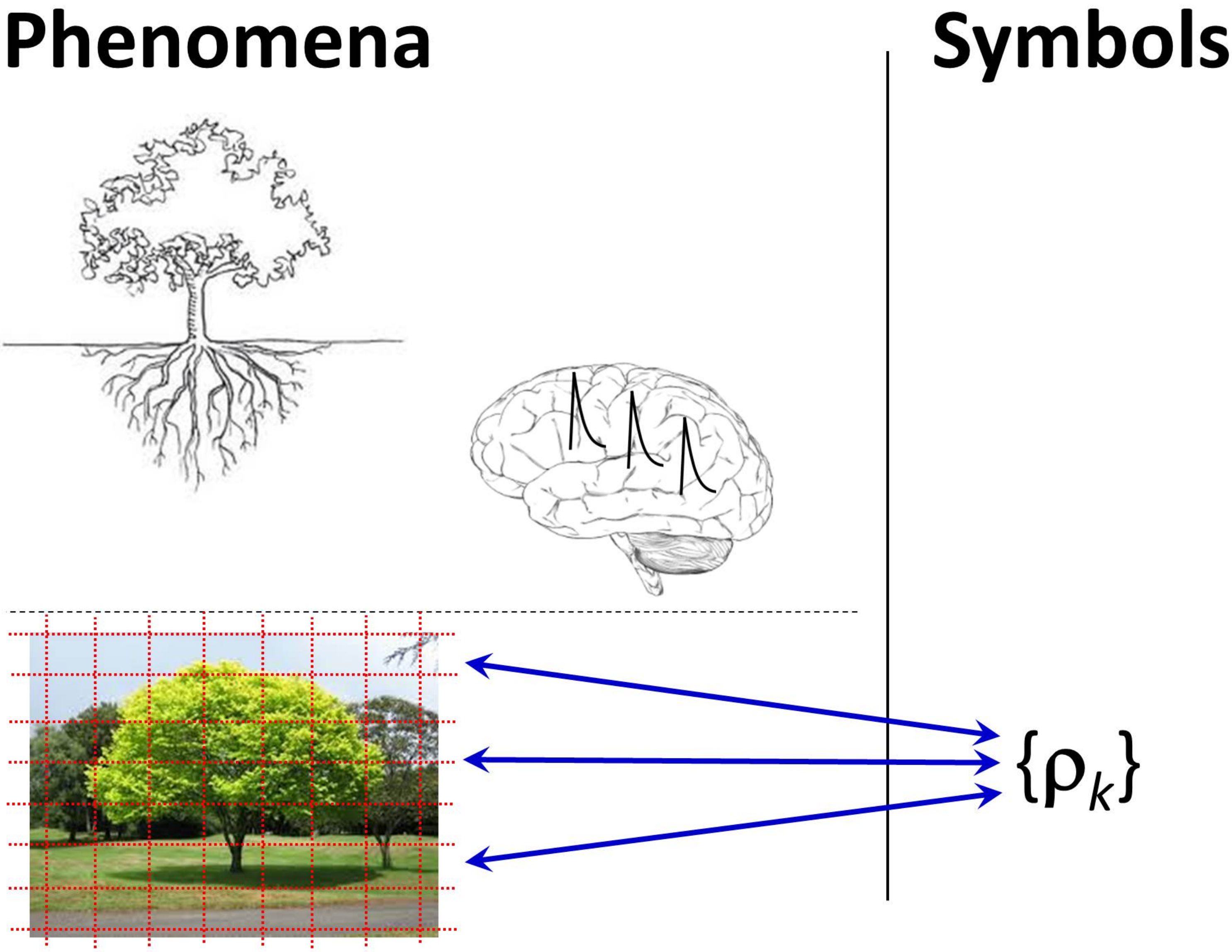

Figure 10.JPEG

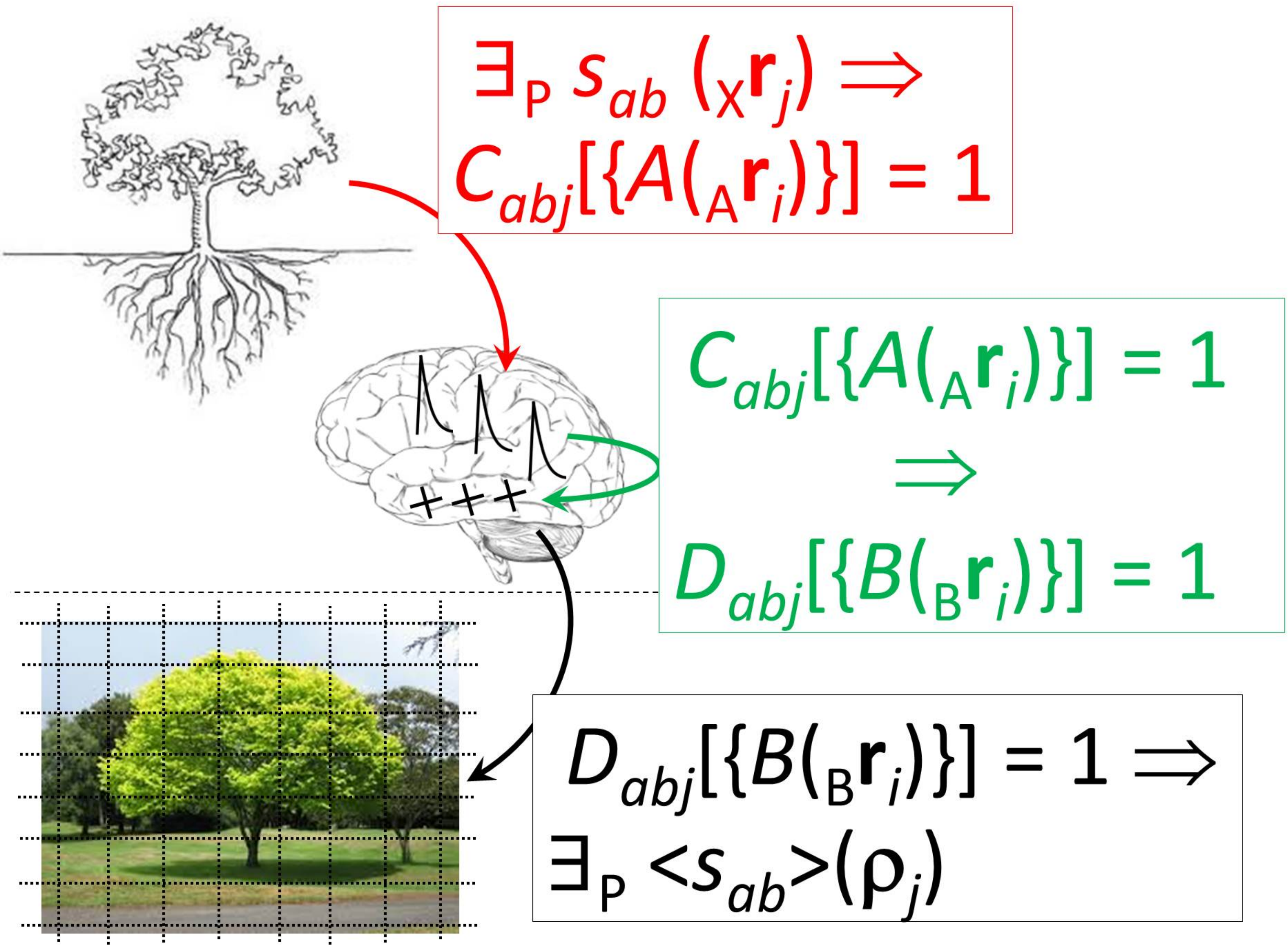

Figure 11.JPEG

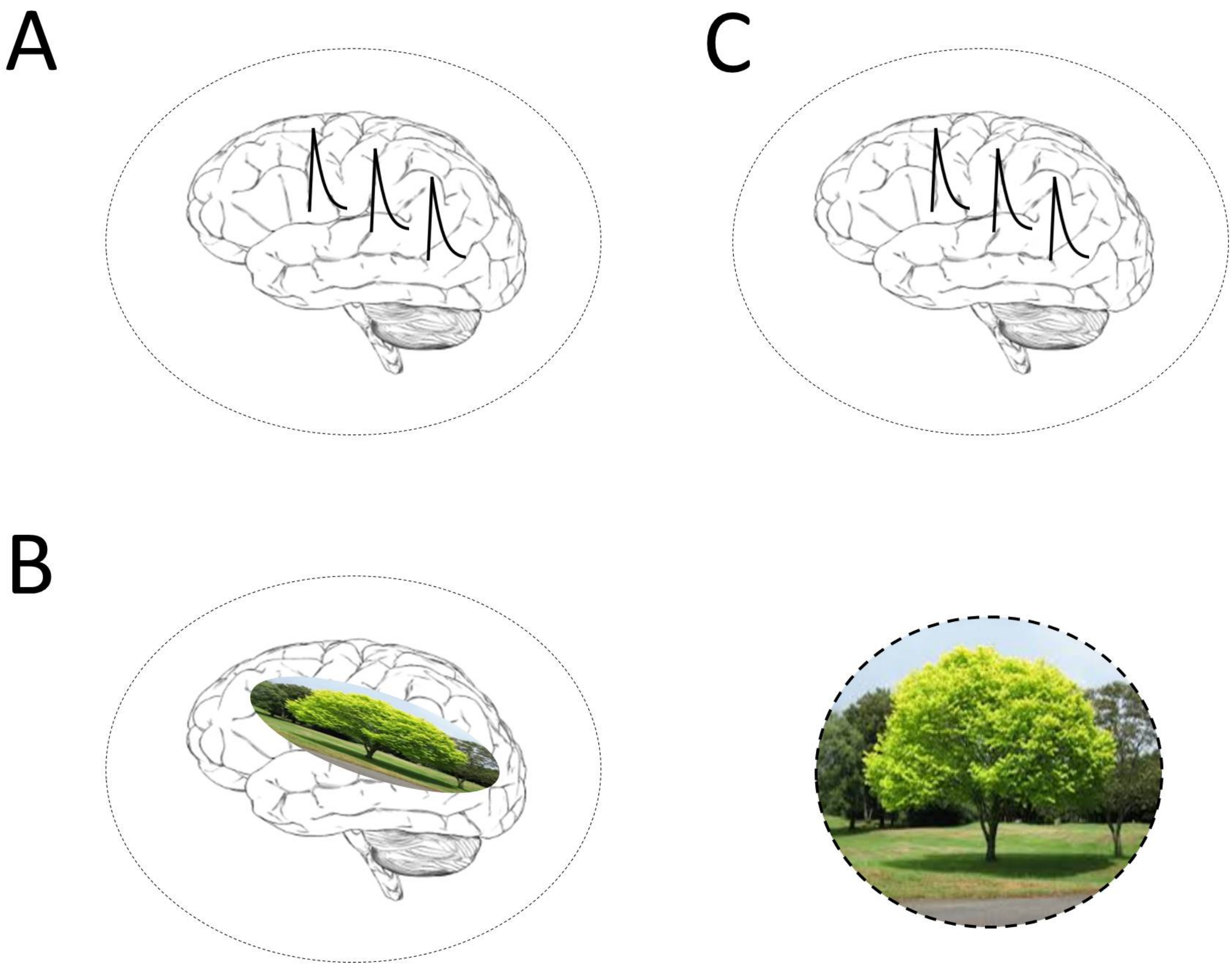

Figure 12.JPEG