The Structure of Phenomenal Consciousness

Jon Opie and Gerard O'Brien
Discipline of Philosophy
University of Adelaide
South Australia 5005

Final Draft @ March 2012

Abstract

Philosophers have largely rejected *strong reductionism*: the view that explanations in the special sciences will ultimately be replaced by explanations couched in the language of microphysics. Critical to this reaction is the recognition that high-level kinds are *multiply realizable,* and thus not (type) identifiable with their microphysical constituents. As a result, it is widely held that the special sciences trade in *functional* explanations, and that the entities they appeal to have *functional essences*. However, this view is not borne out by careful examination of scientific practice. The explanatory strategy of the special sciences is neither reductionist *nor* functionalist, but *mechanistic*. Mechanistic explanations appeal to active entities organized so as to produce the target phenomena. This strategy privileges neither constituents, nor their causal roles, but instead emphasises the various kinds of organization—spatial, temporal, and structural—responsible for the behaviour of material systems. Cognitive scientists regularly apply this approach to perceptual and cognitive phenomena (Bechtel 2008). We claim that phenomenal consciousness will also succumb to mechanistic explanation: it will turn out to be the activity of specific neural mechanisms in the brain. In this chapter we explore the implications of this perspective for the ontology of consciousness, arguing that it has a complex *structural* essence.

1. Introduction

Among all the mysteries of the modern era, the nature of phenomenal consciousness is one of the most profound. Many philosophers regard the problem as uniquely difficult: either an ultimate mystery (McGinn 1991), or requiring explanatory principles not yet provided by the sciences (Nagel 1974, Chalmers 1996). We think this gloomy prognosis is a mistake; the result of taking a too narrow view of the nature and power of scientific explanation. Our aim here is to reflect on recent progress in cognitive neuroscience and the philosophy of science, and show how these disciplines jointly promise to illuminate the deep nature of consciousness.

The question before us — What kind of thing is phenomenal consciousness? — is an ontological question. Ontology is the study of what there is. Although sometimes regarded as an arcane pursuit suited only to philosophers, there is a very real and intimate connection between scientific explanation and ontology. To begin with, the form of our explanations is influenced by the nature of the things we seek to understand. Very different mathematical and conceptual resources are required to explain, say, fundamental *physical* processes (such as quantum tunnelling), and *biological* processes (such as mitosis). Secondly, and more importantly for our purposes, we look to our best scientific explanations for guidance as to the essential nature of things.

Philosophers of science generally acknowledge just two kinds of material essence[1]: the *microphysical* and the *functional*. A focus on microphysics issues from the view that the special sciences have a common grounding in subatomic physics. Explanation is here conceived as an iterated series of reductions that will ultimately permit us to comprehend the behaviour of any system, be it chemical, neural, social etc., in terms of processes occurring at the subatomic level. From this point of view, the essence of any material system is its *composition*, that is, its analysis as a

[1] The "essence" or essential nature of something is that which makes it the kind of thing it is. So, for example, the essence of combustion is that it involves a sequence of exothermic chemical reactions between a fuel and an oxidant. By "*material* essence" we mean to distinguish our target from, e.g., logical and conceptual systems, whose essence is arguably non-material. The viability of essentialism is a matter of some controversy among philosophers (see, e.g., Robertson 2008), but we will take it for granted that it provides a useful way to think about ontology.

collection of subatomic constituents. A focus on functional analysis issues from the multiple realizability[2] arguments that dominated discussions in the philosophy of science during the 1970s and 1980s. These arguments led many philosophers to conclude that the special sciences are not reducible to microphysics, because the natural kinds of the special sciences are not (type) identifiable with their subatomic constituents – they have *functional*, rather than microphysical essences.

However, philosophers have recently been paying closer attention to the practice of the special sciences and to the nature of the explanations they generate. What is apparent is that the explanatory strategy of the special sciences is neither functionalist nor strongly reductionist. It is *mechanistic*. Rather than seeking to understand phenomena by way of decomposition into their subatomic parts, the special sciences investigate the manner in which the behaviour of material systems is constituted by relationships among their components. The focus here is on *organization* rather than (microphysical) composition. Nor does this mechanistic strategy appeal to functional essences, since the types of organization at issue are not limited to the causal analyses associated with functional explanation.

Philosophy of science has yet to fully appreciate the ontological significance of this mechanistic strategy. We will argue that the success of mechanistic explanation requires the introduction of a new ontological category. Phenomena produced by mechanisms are best understood as possessing *structural* essences. This result has important implications for our understanding of phenomenal consciousness. Conscious experiences, like many other natural kinds, are not identifiable with their microphysical parts. But neither can they be properly characterized in a way that ignores the material properties of their realising mechanisms. Conscious experiences are emergent phenomena that are metaphysically dependent on their material components, without being identical to them.

Our task in this chapter will be to defend this set of claims. We begin by offering a brief history of the philosophy of science responsible for generating the dichotomy between microphysical and functionalist metaphysics. We then consider at greater length the mechanistic model of explanation. This much is merely descriptive. The chapter turns polemical when we consider the metaphysical implications of this model of explanation, and their application to the case of consciousness. Our conclusion is that conscious experiences, from an ontological perspective, are emergent structures.

2. Microphysicalism versus Functionalism

Throughout the better part of the twentieth century the orthodox account of scientific explanation was one that focused on the "laws of nature". To explain a phenomenon was to derive it from statements describing laws and boundary conditions. This view, most clearly expounded in the work of Hempel (1966), is known as the Deductive Nomological (DN) model, because it treats explanation as the logical derivation of explananda from laws (Greek: nomoi). For example, to explain why the pressure on the walls of a gas-filled piston doubles when its volume is halved, we invoke Boyle's law. This law states that the product of pressure and volume for an ideal gas at a fixed temperature is constant ($PV = c$, where P is pressure and V is volume). The measured change in pressure can be derived using this law, with the change in volume and the other criteria operating as boundary conditions.

What distinguishes one scientific discipline from another, on the DN model, is the theoretical vocabulary, ontology, and proprietary laws that apply to their respective domains. The various gas laws thus serve to pick out the discipline of thermodynamics, which (among other things) deals with gases conceived as fluids with a characteristic set of macroscopic properties. Kinetic theory, by contrast, deals with the behaviour of the microscopic constituents of matter, treated either as classical particles governed by the Newtonian laws of motion, or quantum systems governed by the laws of quantum mechanics. Laws themselves stand in need of explanation, but this is typically achieved using the resources of a more fundamental discipline. To explain Boyle's gas law, for example, we invoke mechanical and statistical principles which do not belong to classical thermodynamics, and which are arguably more wide-ranging. If we can derive the laws of one discipline from the those of another, without remainder, then the former is said to *reduce to* the latter.

With this view of explanation and reduction in mind, Oppenheim and Putnam (1958) developed an influential scheme for organizing natural objects. They divide nature into the following six levels, where each level comprises a set of entities with a characteristic scale, and subject to the laws of an attendant discipline: Elementary Particles (Particle Physics), Atoms (Atomic Physics), Molecules (Chemistry), Cells (Cell Biology), Organisms (Biology), Society (Social

---

[2] A property or kind is "multiply realizable" if its different instances need not have a common set of building blocks. For example, a wing can be composed of bone, sinew and feathers; balsa and canvas; or aluminium and steel.

Science). A principal motivation for this scheme is to provide a basis for the putative unity of science, a unity constituted by reductive relationships among levels. The idea is that social science will eventually reduce to biology (presumably via such intermediate special sciences as psychology and neuroscience), biology to cell biology, cell biology to chemistry, and so on. DN reduction is transitive in this scheme: if cell biology reduces to chemistry, and chemistry to atomic physics, then cell biology also reduces to atomic physics because the derivation of cell biology piggy-backs on the derivation of chemistry, allowing one to explain cellular phenomena in atomic terms. Were this sequence of reductions to be completed, all of science would be unified by its grounding in subatomic physics, which would provide a common starting point for the derivation of phenomena at all levels.

So the DN model of explanation goes hand in hand with *strong reductionism*: the view that the special sciences will eventually be replaced by explanations couched in the terminology of subatomic physics. This view results from privileging the ontology of microphysics over all others. We might agree that all phenomena in the universe ultimately depend on the elementary parts of matter, in the sense that were the latter to spontaneously disappear, so would the former. But strong reductionism suggests a further, more radical claim, namely, that the universe and all the systems it contains is *nothing but* a vast congeries of elementary particles. From this perspective, conscious experience, like every other material phenomenon, is nothing more than a (mind-numbingly) complex conjunction of sub-atomic particles behaving in accordance with the fundamental laws of physics.

During the 1970s and 1980s, this reductionist conception of the unity of science started to come unstuck, in large part because of so-called *multiple realizability* arguments. In an important early paper, Fodor (1974) argued that the special sciences don't reduce to (micro)physics, because the properties over which their laws are defined don't have unique decompositions into the properties of elementary particles. While individual cells, tectonic plates, airfoils, galaxies, and so on, are no doubt composed of fundamental particles, the property of being a cell, a tectonic plate, an airfoil, or a galaxy is realized by such a heterogeneous and unsystematic set of microphysical properties that, in principle as well as in practice, no reduction of the laws of the special sciences is possible. This analysis precipitated an avalanche of further reflections on the significance of multiple realizability, and resulted in widespread abandonment of the reductionist program.

The ontological consequences of the demise of strong reductionism were immediate and profound. Most theorists embraced the view that the phenomena of the special sciences have *functional*, rather than microphysical, essences. This is the basis of the *functionalist metaphysics* that dominates contemporary philosophy, just as microphysicalism did during the first half of the twentieth century. According to functionalism, special science kinds belong to the abstract domain of *causal organization*. This view appears to follow immediately from the existence of multiple realizability. If there is nothing materially in common to the entities targeted by some special science, what *could* unite them apart from causal organisation? Airfoils, for example, are certainly composed of subatomic particles. However, when it comes to aerodynamics, it is not the composition of an airfoil that matters, but rather the pattern of causal relations into which it enters, or so a functionalist would argue.

Functionalism has a number of significant implications for the relationship between the sciences. Most significantly, it licenses the *autonomy* of the special sciences. Because each science operates with its own proprietary level of causal organisation, even cognate disciplines will have little to say to one another about the fundamental nature of their respective kinds. It was functionalist metaphysics that motivated strong claims about the independence of cognitive science from neuroscience late in the twentieth century. Although neuroscience undoubtedly has something interesting to say about the substrate of our own cognitive processes, the essential nature of cognition is not illuminated by analysing the material properties of the brain.

Although multiple realizability arguments have been almost universally thought to undermine strong reductionism, a number of their proponents, including Fodor, did not take them to undermine the DN model of explanation. However, at more or less the same time as these arguments were proving destructive of reductionism, other philosophers were applying pressure more directly on the DN model. In an influential paper, Salmon (1978) argued that scientists generally seek to discover the *causes* of some range of phenomena, rather than subsume them under laws. To explain variations in the pressure on the walls of a piston one describes: i) the system involved — a fixed quantity of gas made up of a huge number of fast-moving particles, a closed but expandable container with rigid walls; and ii) the way the parts of the system produce the phenomenon — pressure variations are caused by changes in the mean rate of particle-wall impacts as a result of changes in the size of the container. In other words, one describes the *mechanism* that produces the explanandum phenomenon.

3. The Mechanistic Model of Explanation

The mechanistic model of explanation has a venerable history, dating back to Democritus and the Greek atomists. During the early modern era it was championed by the likes of Galileo Galilei, Rene Descartes, and Robert Boyle (Bechtel 2006, pp.20-22). Recent work in this tradition has focused on providing a general account of the nature of mechanisms and their role in scientific explanation (Bechtel 2006, 2008; Bunge 1997; Craver 2001, 2007; Glennan 2002; Machamer, Darden & Craver 2000; Woodward 1989).

Our everyday conceptions of mechanism, which are informed by experience with relatively simple devices such as clocks and corkscrews, are potentially at odds with scientific usage. Such artefacts certainly embody mechanisms, but they are poor models for the kinds of processes scientists typically invoke to explain natural phenomena. Natural mechanisms — such as thermal conduction, protein synthesis, bacterial conjugation, perspiration, sexual selection, gravitation, and so on — need not have rigid parts, nor must they be linear, denumerable, or easily understood. Indeed, many natural phenomena result from complex, non-linear processes that defy our best attempts at analysis.

In the broadest terms, a mechanism is a process in a material system that produces (or prevents) some change or brings something into being.[3] Material systems themselves are sometimes referred to as "mechanisms" but it is more consistent with scientific usage to identify mechanisms with productive processes. Bechtel stipulates that "[a] mechanism is a structure performing a function in virtue of its component parts, its component operations, and their organization" (2006, p.26). Likewise, Craver regards a mechanism as "a set of entities and activities organized such that they exhibit the phenomenon to be explained" (2007, p.5). The emphasis here is on the way operations or activities are organized to produce a specific outcome. To specify a mechanism one must identify: i) the relevant parts of the system, ii) the activities of those parts, and iii) how the organization of those parts and their activities gives rise to the phenomenon of interest. For example, the pumping of the heart depends on certain of its parts (ventricles, atria, valves), their activities (contraction and relaxation, opening and closing), and their spatial, temporal and causal relations (valves connect chambers and vessels, atria and ventricles contract and valves open in a specific sequence).

According to the mechanistic account of explanation, to explain some behaviour or capacity Φ of a material system S is to identify and describe a process M in S that produces Φ. If S is an open system, M will be subject to outside influences, so fully specifying Φ will involve specifying the environmental conditions that bear on M. For example, blood is part of the heart's environment, and the condition of the blood has an impact on the heart's activity. If we aim to explain the pumping of the heart, we must first establish how variation in the properties of the blood (pH, pO2, viscosity, etc.) affects this behaviour. Thus, mechanistic explanation almost always encompasses at least two distinct levels: i) the level of the system S and its environment, including any relevant containing systems; and ii) the level of the active parts of S (Bechtel 2008, p.148). Since mechanistic explanation iterates, deeper explanations will reveal further levels of organization in the structure of the parts of S.

Bechtel (2008) and Craver (2007) argue that material organization is the basis of this level hierarchy. As remarked, every mechanistic explanation involves at least two levels of organization, and mechanistic explanations typically also refer to the organization of one or more subsystems of the primary system. What belong at each level are the active entities whose organization produces the explanandum phenomenon at the next level. Thus, atria, ventricles and valves are at one level, because their organized activity constitutes the pumping of the heart; the heart, blood and vessels a next higher level, because they collectively act to circulate nutrients, hormones and gases around the body. A mechanism at one level is *composed of* entities at the next lowest level of organization. Those entities are themselves composed of entities at a yet lower level, and so on. Composition implies spatial and temporal containment, that is, the parts and activities of a mechanism cannot exceed the size and duration of the whole process. However, it is important to recognise that a mechanism is no mere bag of parts. A mechanism is a *physical gestalt* (Köhler 1920); a material whole that is constituted by the organization of its parts and their activities.

There are a number of significant respects in which this mechanistic approach to levels differs from Oppenheim and Putnam's scheme. First, mechanistic levels are always local (Craver 2007, pp.192-3). They do not comprise a single, global hierarchy, because they are always defined relative to the current explanatory target. Second, the order associated with mechanistic levels is partial, not total (Bechtel 2008, p.147; Craver 2007, p.191). Some things are

---

[3] This formulation is adapted from Bunge (1997). A material system is a bundle of real things that behave in some respects as a unit by virtue of their interactions or bonding. Atoms, crystals, synapses, transistors, cells, organisms, families, firms and galaxies are material systems, which are to be contrasted with conceptual systems, such as theories and classifications. (ibid, p.415)

related by level membership, others are not. Even though the heart is composed of cells, osteocytes (bone cells) do not appear at a lower level than the heart in the mechanism of circulation. Third, entities of a given kind can appear at more than one level in the same mechanism. For example, some sensory neurons detect acidity via proton-gated ion channels, which open a pore in the cell membrane, thereby inducing a sensory signal (Waldmann et al, 1997). Free protons are at the same level as ion channels in this context, because the two must interact for acid-sensing to occur. However, since protons are a basic building block of matter, protons are also constituents of the ion channels themselves. Thus protons appear on at least two organizational levels in this mechanism.[4]

The mechanistic approach to levels is at odds with some very entrenched intuitions. It is by now second-nature to associate levels with well-defined types, each having a fixed place in a monolithic, global hierarchy. The simplicity of this picture is one of its chief attractions, but also its principal flaw, because it fails to reflect the complexities of scientific practice. Most disciplines pay little attention to boundaries defined by scale or type, seeking out whichever entities, activities and forms of organization illuminate their explanatory targets. Indeed, it is reasonable to ask why anyone would expect a global level hierarchy to make sense of the very different explanatory demands of, say, cell biology and plasma physics. Both disciplines investigate processes whose ultimate constituents are protons, neutrons and electrons, but organized in such vastly different ways that the two domains essentially have no structures in common at any scale.[5]

There is both reduction and emergence here. Mechanisms emerge from the organized activity of lower-level entities, and have effects that their constituents in isolation lack. The existence of this kind of novelty is not miraculous, although it can be difficult to understand, especially when the organization involved is complex or non-linear. Mechanistic explanation is reductive in the sense that it reveals the connection between phenomena and the lower level entities that produce them. Yet it doesn't thereby eliminate high-level phenomena, which depend not only on the constituents of their mechanisms, but also on the way those constituents are organized. Nor does mechanistic explanation eliminate disciplines, in the sense of making them redundant. Disciplines span levels, and must cooperate to discover the multi-level mechanisms of complex phenomena. The unity of science, such as it is, does not depend on reductive relationships among disciplines, but on the shared ambition to reveal the hidden structure of nature.

4. The Metaphysics of Mechanistic Explanation

The outline of a mechanistic approach to explanation has been well delineated by a number of philosophers. What of its ontological entailments? Given its rejection of a hierarchical conception of the unity of science, and of strong reductionism, the mechanistic model clearly doesn't comport well with microphysicalism. This, we think, has prompted most philosophers to assume that it leads inevitably to functionalism. But this assumption sits awkwardly with several aspects of the mechanistic story.

One thing that distinguishes mechanistic explanation from abstract functional analysis (Cummins 1975) is a focus on the material basis of systemic behaviours. A functional analysis decomposes an overall system capacity into a set of sub-capacities, but is typically silent on how those sub-capacities are realized. By contrast, a mechanistic explanation reveals how some phenomenon depends on the constitution and organization of particular material entities. Organization has spatial, temporal, and hierarchical dimensions. The parts of a mechanism have characteristic structures, positions and arrangements. Their activities occur with specific timings, rates and durations, and in sequences or cycles which may incorporate feedback and other kinds of orchestration. Organization also has a hierarchical dimension because mechanisms typically contribute to the behaviour of superordinate systems, and are composed of subordinate systems with structure of their own. (Bechtel 2008, pp.10-17; Craver 2001)

This emphasis on material structure suggests that the mechanistic model of explanation does not entail a functionalist metaphysics. But how can this disconnect between mechanism and functionalism be made consistent with multiple realizability? Let us consider one of Fodor's favourite examples: the humble airfoil (1989, pp.61-2). It is certainly true, as Fodor avers, that airfoils can be constructed from a variety of materials—wood, sheet metal, fiberglass, even sheets of canvas—and hence that they need have nothing in common with respect to their composition. It is also true that all airfoils, regardless of composition, have a common causal role: they all bring about lift by causing air to flow more rapidly across one surface than another. What is not true is that this functional property is the *only* thing airfoils have in

---

[4] See Bechtel (2008, p.147) for a similar example.

[5] A plasma is a high-temperature gas of unbound protons, neutrons and electrons, which do not combine to form atoms because of their high kinetic energy.

common. An airfoil generates lift when it moves through the air because its upper surface has *greater curvature* than its lower surface. Airfoils are thus distinguished from other kinds of things not merely by their causal role, but also by their *shape*.[6] And that property is multiply realizable: an airfoil can be constructed from any material rigid enough to maintain the differential curvature of upper and lower surfaces.

If we think more generally about shape, it is clear that an object's shape is not fixed by its subatomic composition. But neither is it determined by its causal role or causal organization (even if the causal relations into which an object enters are sometimes determined by its shape). Metaphysically speaking, shape lies somewhere between these two extremes. Like causal or functional role, shape abstracts away from the microphysical composition of an object. But unlike causal role or causal organization, shape is an *intrinsic material* property.

We need a general term to describe such abstract material properties. We will call them *structural* properties. The functional properties of an airfoil are clearly multiply realizable, but so are its structural properties. The question that arises, of course, is whether airfoils are better classified as possessing a functional or a structural essence. Is it structure or function that determines membership of this kind? In answer, we note that the class of objects that can generate lift is much larger than the class of airfoils — think of jet engines, for example. What distinguishes airfoils from other lift-generating mechanisms is their characteristic curvature. Furthermore, it is this structural property which explains the functional characteristics of an airfoil, not the converse. This suggests that structural properties are more fundamental than functional ones, and that any classification into kinds should privilege structure over function.

On the basis of this tripartite distinction between microphysical, structural, and functional properties, we would argue that the special sciences are founded principally on the metaphysical emergence of structural kinds. Many of the examples philosophers use to illustrate multiple realizability turn out on closer inspection to highlight the significance of structure for natural kinds. Cells, for example, as the nexus of a host of biological processes, and the product of a long and complex causal history, are often treated as the quintessential functional kind. But there are many respects in which a cell is fundamentally reliant on structure: a cell is an entity whose very existence depends on the presence of a semi-permeable boundary, the cell wall, that marks the distinction between the cell's internal components and the environment in which it makes a living. This boundary ensures that the internal milieu of a cell is maintained in its highly organized, far-from-equilibrium condition, in the face of enormous variability in the cell's surroundings. And it hardly needs saying that a cell wall is a (highly specialized) material structure. So in this case, function again follows form, and structure appears to be essential to the identity of the kind in question.

Perhaps even more importantly, many special science kinds which have not been considered multiply realizable are best understood as possessing structural rather than microphysical essences. Ever since the discovery of isomerism by Liebig and Wöhler in the 1820s it has been known that chemical compounds cannot be identified with their atomic constituents. The mercury salts of cyanate and fulminate, for example, are composed of mercury, carbon, nitrogen and oxygen atoms in identical proportions, yet these two compounds have very different chemical properties. Mercury fulminate is a highly unstable substance that explodes under light impacts, friction or heating. It was used extensively during the industrial revolution as a detonator (Kurzer 2000). Mercury cyanate, by contrast, is thermally and mechanically stable. Why do these two substances behave so differently, despite their common atomic basis? It turns out that materials with identical atomic constituents can have distinct molecular structures. The structure of the cyanate ion is O-C≡N whereas in the fulminate ion the nitrogen and carbon atoms change positions: O-N≡C. What distinguishes the two mercury salts is thus not a matter of the nature or proportion of their constituents, but differences in their bonding patterns.

Here we have the converse of multiple realizability: rather than a single higher level kind with different realizers, we have distinct special science kinds constructed from the *same* constituents. Let's call this a case of *multiple composability*. A property or kind is multiply realizable if its instances don't share a common set of building blocks. A set of parts are multiply composable if they can be combined and organized so as to produce numerous distinct higher-level kinds.

Multiple realizability undermines microphysicalism, you will recall, because special science kinds are typically realized by such a heterogeneous set of microphysical properties. Multiple composability leads to the same conclusion via a different route. To make sense of the very different physical and chemical properties of the cyanate and fulminate ions

---

[6] The lift generated by an airfoil, such as a wing, doesn't just depend on its shape, but also on its surface area and angle of attack. Nonetheless, the term "airfoil" is usually reserved for objects with differential curvature of their two principal surfaces.

it does not suffice to identify their constituents. The mechanism that explains the explosive properties of mercury fulminate, for example, depends not only on its lower-level atomic components, but also on the manner in which they are combined, in particular, the triple-bond on the carbon atom causes it to be negatively charged and thereby renders the ion thermodynamically unstable.[7] Furthermore, since an explosion is a macroscopic phenomenon involving vast numbers of ions (or molecules), a full account of the detonation of mercury fulminate will describe chemical and mechanical processes occurring at the level of its crystalline structure (see, e.g., Brown & Swallowe 1981, Faust 1989). Microphysicalism is inadequate to this kind of explanation, because it downplays or ignores the ontic significance of emergent structure: the various levels of organization that occur in any real-world system. By the same token, a functionalist metaphysics fails to grapple with the consequences of multiple composability, because the kinds it envisages are so abstract as to render invisible the very processes that distinguish, say, fulminates from cyanates.

Multiple composability is ubiquitous in the sciences. Just as one set of atomic constituents can compose distinct chemical compounds, the same cell types can compose different biological organisms, the same neural cells different neural architectures, the same neural networks different cognitive mechanisms, and so on. Composition alone is not of the essence for the kinds of the special sciences, nor, it seems, is causal organization. Material structure trumps function in mechanistic explanation, hence special science kinds, whenever they feature in such explanations, are best conceived as having structural essences.

So much for the philosophy of science. It is now time to apply the mechanistic approach to the explanation of phenomenal consciousness.

5. The Ontology of Consciousness

In the foregoing we sought to tease out the general metaphysical implications of mechanistic explanation in the special sciences. Our next task is to apply the lessons learned there to the specific case of consciousness. We saw earlier that according to the mechanistic model, to explain some phenomenon Φ in a material system S one must identify: i) the relevant parts of S, ii) the activities of those parts, and iii) how those parts and their activities are organized so as to produce Φ. As other contributions to this volume demonstrate, contemporary neuroscience has made some progress in identifying the parts and activities of the brain involved in producing consciousness. What remains to be achieved is some consensus on how the organization of the brain gives rise to this phenomenon.[8] In this section we will provide a brief overview of the account of consciousness that is emerging from neuroscience, and consider what it implies about the ontology of consciousness.

Although there are some dissenters, the neuroscience community is converging on the view that consciousness is produced by a neural system that incorporates the cerebral cortex and thalamus, together with their dense bidirectional connections. Rather than a single, anatomically discrete consciousness-making centre in the brain, consciousness seems to depend on activity in both cortical and subcortical structures. The evidence for this hypothesis comes from a variety of sources, but especially from an examination of the neural activity associated with sleep, anaesthesia, coma and epileptic seizures (Revonsuo 2010, pp.159-64). There is a significant amount of activity in this system even during NREM sleep, anaesthesia, and generalized epileptic seizures, when consciousness is largely absent. So it is reasonable to conclude that "some additional dynamic feature of neural activity must be present to generate conscious content" (Tononi & Koch 2008, p.248).

Lamme (2000, 2004) argues that conscious experience is the result of *recurrent* neural activity. The processing of visual information in the cortex is marked by two distinct phases: a rapid feedforward sweep that reaches V1 about 40ms after stimulus onset, and recurrent processes that involve feedback between early and late visual areas. The feedforward sweep influences activity in all extra-striate visual areas by around 80ms, but recurrent interaction doesn't arise until 100ms after stimulus onset. Lamme contends that the feedforward sweep, despite its ability to influence

---

[7] Fulminate is actually a resonance hybrid of two principal Lewis structures, O-N≡C and O=N=C, both of which result in a negative formal charge on carbon and a positive formal charge on nitrogen. It is this energetically unfavourable electron configuration that makes fulminate so unstable.

[8] Admittedly, it remains a tough problem to discern which of the brain's activities *constitute* consciousness, as opposed to being merely correlated with it (Miller 2007). To fully address this problem we need a detailed mathematical account of the structure of consciousness, and a more sophisticated understanding of the dynamics of neural activity. These will ultimately provide the basis for an exhaustive neuro-phenomenology, which will solve the constitution problem and also ease the minds of those troubled by Chalmers' "hard problem" (1996).

activity throughout the brain, is not sufficient for visual experience. Instead, visual consciousness is the result of recurrent interactions between early and late visual areas.[9] Recurrent activity depends on both horizontal intra-area connections, and feedback-feedforward circuits between areas, so there is a tight coupling in Lamme's account between the anatomical and physiological forms of organization that are required for (visual) consciousness.

Not every neuroscientific account of consciousness addresses the mechanism of consciousness head-on. Tononi (2004, 2008), for example, argues that what renders a cognitive system conscious is its capacity to integrate information. This capacity is not an all-or-nothing affair, but depends on the size and complexity of the system, its current level of activity, and the nature of the input it receives. Stated in such general terms this might not sound like a hypothesis with any implications for the mechanism of consciousness. However, since the brain's capacity to integrate information depends on the ability of neural networks to influence each other, either directly or indirectly, Tononi's hypothesis suggests that this mechanism involves high levels of reciprocal connectivity and physiological processes that enhance inter-areal communication, e.g., synchronized neural firing (Singer 1999).

Llinas et al (1998), Newman (1995) and Revonsuo (2006) argue for the importance of thalamocortical interaction in the production of conscious states. These interactions are supported by both *specific* and *non-specific* thalamocortical loops. Specific thalamic nuclei, such as the lateral geniculate nucleus, are reciprocally connected to specialized cortical areas (V1 in the case of the LGN), whereas non-specific nuclei, such as the pulvinar, connect to multiple cortical areas, relaying information between first-order and higher-order cortical areas. There is increasing evidence that these two systems are involved in synchronizing cortical activity, binding anatomically distinct regions into transient neuronal assemblies (Varela et al 2001, Singer 2007). Llinas et al suggest that both systems are necessary for consciousness, the former binding specific sensory and motor processes into locally synchronous patches of gamma-band (20-50Hz) activity, the latter providing a global, unifying context by coincidence detection right across the cortex.[10] On this picture, conscious experience is a globally unified pattern of recurrent bioelectrical activity operating with a characteristic temporal dynamics.

A number of issues remain open here. It is unclear whether thalamocortical synchrony is sufficient to generate conscious experiences, or if other features of neural activity, such as a *recurrence* (Lamme 2004) or *stability* (Tononi & Koch 2008, p.248-50, O'Brien & Opie 1999) might also be essential. There is also disagreement about the amount of cortical tissue required to generate a conscious state. Zeki and Bartels (1999) argue that consciousness-making in the brain is highly localized and modular, such that, for example, each distinct processing module in the visual cortex generates an element of visual experience – a "microconsciousness". Whether or not one believes that conscious states can occur at this scale, the extent to which global consciousness is dissociable into independent phenomenal parts remains an open question (O'Brien & Opie 1998, 2000).

Despite these points of difference, recent neuroscientific accounts of consciousness have two important features in common: i) they are all mechanistic theories, and ii) they all equate conscious experiences with specific kinds of activity in the neural networks of the thalamocortical system. Our focus here is the metaphysical implications of this consensus. Neuroscientists sometimes claim that their work only reveals the *neural correlates* of consciousness. Such caution is commendable on empirical grounds. It is a significant methodological challenge to demonstrate conclusively that one thing causes another, or that some emergent phenomenon is constituted by the activity of systems at a lower level of organization.[11] But whether they realise it or not, theorists who explore the mechanisms of consciousness are going beyond mere correlation (Revonsuo 2006, pp.293-303). Each of the hypotheses surveyed above is an *identity claim* — an assertion to the effect that conscious experience (or a particular class of conscious experiences) is *none*

---

[9] There are many pieces of evidence for this claim. For example, feedforward activation of neurons can be recorded in anesthetized animals with receptive field properties that hardly differ from those in the awake state, whereas recurrent processing is largely suppressed or absent.

[10] When a neural network is locked into gamma-band oscillations, this doesn't imply that its constituent neurons are all firing at the same rate, but rather that the network as a whole is dominated by activity in this frequency range. Gamma-band activity is an emergent statistical feature of a neural population. Fries et al (2007) have suggested that sensory and motor information might be encoded in the temporal offset of spikes relative to the gamma-band "metronome".

[11] See Craver 2007 for a detailed discussion of this point.

other than a particular kind of activity in the thalamocortical system of the brain. At one blow these proposals dispense with dualism[12] and offer us a variety of mechanisms whereby the brain might generate conscious experiences.

An identity claim equates one thing with another. The classic philosophical example is the identification of lightning with electrical discharges. Prior to the experiments of Thomas-François D'Alibard and Benjamin Franklin in 1752, the nature of lightning was still a matter of conjecture. Franklin found that lightning could be communicated along the string of a kite or a metal rod during a storm, giving rise to discharges which behave in every respect like discharges of terrestrial origin (Priestley 1775, p.216-8). He thereby provided the first evidence that these two seemingly distinct phenomena have the same essential nature: lightning just is an (atmospheric) electrical discharge. This is identity "in the strict sense" (Smart 1959, p.145), in which a single kind of thing is known under two descriptions. Likewise, when a neuroscientist proposes a mechanism for consciousness, this is an attempt to identify conscious experiences with particular processes in the brain, processes that are subject to both first-person and third-person descriptions, but which nonetheless constitute a unitary phenomenon.

Existing attempts to identify conscious experiences with a particular class of neural mechanisms already tell us a great deal about the metaphysics of consciousness. To begin with, they rule out microphysicalism, which identifies consciousness with the activity of the subatomic constituents of the brain. We have already rejected this form of reductionism on the grounds that it recognises neither the multiple realizability of special science kinds, nor the multiple composability of their parts. A less extreme view might be that consciousness will ultimately be understood in terms of the *neural* composition of the brain. But again, this approach fails to grapple with multiple composability, and specifically, the discovery that neural systems can give rise to both conscious and unconscious processes.

So we return to the view that consciousness is a phenomenon which depends in some way on the organization of the brain. Most philosophers take this to settle the matter in favour of functionalism. On this view, what neuroscientists are pointing to when they identify particular kinds of neural activity with conscious experience are objects or events that play a particular role in the causal organization of the brain. And it is that causal role, that specific set of causal connections with other processes and events, that is the essence of consciousness. In other words, what makes a pattern of neural activity a conscious experience has nothing (directly) to do with its intrinsic material properties, and everything to do with its causal role.

Although it appears unobjectionable when stated in such abstract terms, this position has the unpalatable consequence of rendering conscious experience causally inert. Functionalism is committed to the view that consciousness is a mere by-product of brain activity with no impact on the subsequent behaviour of the system, rather like the static produced by an AM radio. This result is perhaps unexpected, given functionalism's focus on the causal organization of the brain, but it follows directly from the fact that functionalism identifies consciousness with a causal *role*. A causal role has *no effect on anything*, rather, it is the material state or process that fills that role which is causally efficacious. Functionalism explicitly separates role from realizer, and identifies consciousness with the former, not the latter. In slightly different language, functionalism identifies consciousness with the *doing*, not the *doer*.

But of course, we are not stuck with functionalism. Our discussion in the previous section suggests an alternative metaphysics, one which, like functionalism, focuses on the organization of the brain, but which better reflects the mechanistic style of explanation employed by neuroscientists. To describe a consciousness-making mechanism is to identify conscious experience with the activity of a particular kind of material system, and show how it emerges from the organization of that system. It would be tempting to suggest, as does Revonsuo (2006), that consciousness is in fact a level of organization in the brain. But this isn't quite right, at least not if one adopts the definition of mechanistic level we gave in Section 3. What belong at each level in a mechanistic hierarchy are the entities whose organization produces the phenomena at the next level, the level at which those parts and their activities constitute an organized whole of some kind. The heart, vessels, blood, and valves of the circulatory system are a level of organization in the mechanism that delivers oxygen and nutrients to the body. But picked out in this way, these entities are merely a collection of parts, not a working mechanism. To produce their characteristic activity, those parts must be organized in the right way. Likewise, a collection of neural systems do not amount to a conscious experience, rather, consciousness is constituted jointly by those systems, their activities, and the manner in which they are organized

[12] The view that phenomenal consciousness depends on laws, principles or properties that are not within the scope of the sciences as we currently conceive them.

(Miller 2007). In other words, consciousness is a certain kind of *material structure*, and the metaphysics appropriate to a mechanistic account of consciousness is *structuralism*.[13]

What does consciousness look like from this perspective? According to structuralism, it is in virtue of possessing certain structural properties that a pattern of neural activity constitutes a conscious experience. Structural properties, like functional properties, are *abstract* in the sense that they may be realized by a variety of substrates. But unlike functional properties, structural properties are characterized by specific types of material organization. An airfoil must have a particular shape; a cell wall, the right topology; a detonator, the right bonding structure, and so on. Two kinds of organization are germane to our understanding of consciousness: i) those that distinguish conscious from unconscious brain activity, and ii) those that distinguish one species of conscious experience from another, e.g., a visual experience from an auditory experience. Most of the proposals described above address the former issue. To finish, we will briefly explore an idea about the kind of material organization that might distinguish one species of experience from another.

It turns out there is a branch of cognitive science that has already been investigating the structure of patterns of neural activity, albeit in a very simplified form: the neurocomputational approach to cognition known as *connectionism*. Connectionist networks are idealised models of real neural networks, which, although unrealistic in certain respects, capture many of the salient features of mammalian cognition (see, e.g., Bechtel & Abrahamsen 2002). A connectionist network consists of a set of interconnected processing units, each of which has an activity level that is communicated to other units via modifiable, weighted connections. Each unit sums the weighted activity it receives, and generates a response that is some threshold function of its current activity and that input. Via this process, a network transforms patterns of activity across its input layer into patterns of output activity. Altering the network's connection weights alters the nature of the mapping between input and output layers. Consequently, a network can be taught to generate a range of target patterns in response to a range of inputs.

For our purposes, one of the most interesting results of connectionist research is that networks with an identical architecture, but with distinct initial (random) assignments of connection weights, can learn to perform the same input-output mapping. Each network does this by way of a unique final configuration of weights. Some theorists see this as a vindication of functionalism: the only thing that such "microphysically" distinct networks appear to have in common is the input-output mapping they perform, that is, a particular set of causal relations between inputs and target outputs.[14] But closer inspection reveals a deeper kind of similarity.

Consider Cottrell's face-recognition network (Churchland 1995, pp.38-55). This network has a three-layer feed-forward architecture: a 64x64 input array, fully connected to a hidden layer of 80 units, which in turn is fully connected to an output layer comprising 8 units. Each unit in the input layer can take on one of 256 distinct activation values, so it is ideal for encoding discretised grey-scale images of faces and other objects. After squashing through the hidden layer these input patterns trigger three units in the output layer that code for face/non-face status and gender of subject, and five which encode arbitrary 5-bit names for each of 11 different individuals. Different realizations of this network architecture, trained on the same data set, have distinct weight values. However, cluster analysis reveals that they have something significant in common: their hidden layer activity is organized in the same way.

This similarity can be visualised in terms of an *activation space*, in which the activity of each hidden unit is represented using a distinct coordinate axis. A pattern of hidden-layer activation corresponds to a point in this space, and physically similar patterns are represented by nearby points (as measured using the standard Cartesian metric). Post-training analysis reveals that each face-recognition network partitions its hidden unit activation space in the same way: into face and non-face regions, male and female sub-regions, and still smaller regions corresponding to the cluster of patterns associated with each subject (see Figure 1). Each network captures the relations among faces that underlie our judgments of gender, identity, etc., in the distance relations among the points in its hidden unit activation space. Faces that are similar are coded by nearby points, dissimilar faces are coded by points that are further apart.

Figure 1 about here.

---

[13] Notice that the structure of consciousness is *dynamic*, like the organization of the homeostatic processes in a cell, rather than *static*, like the macroscopic structure of an airfoil.

[14] In this context, a "microphysical" difference is a difference in connection weights, *not* a difference at the atomic or subatomic scale. Two networks can share a network architecture, i.e., have the same number of units at each layer and an identical set of connections between layers, yet be microphysically distinct in this sense.

So, although it is certainly true that microphysically distinct face-recognition networks have a common causal profile, this functional similarity is grounded in something else: a common partitioning of hidden unit activation space. A partitioning is an abstract yet intrinsic material feature of a network. Since activation patterns are themselves physical objects, the distance relations in the activation space of a network codify *physical* similarities and dissimilarities between those patterns. And when such distance relations are preserved across networks, it marks the fact that despite differences in their connection weights, these networks embody a common *set* of similarity and dissimilarity relations. This is a shared, but abstract property. And it is an abstract property that is grounded in the intrinsic material organization of each individual network. Consequently, when networks partition their hidden unit activation spaces in the same way, what they have in common is a *structural* property, a property, moreover, that underpins and explains their functional similarity.

Despite the idealized nature of the connectionist analysis of neural activity, it may well have something to teach us about the ontology of consciousness. It provides a framework for understanding how patterns of neural activity can possess common structure, in the face of microphysical differences, and how these abstract properties underlie the capacity of distinct networks to enter into similar causal relations. Furthermore, the identification of such structural properties allows for both multiple realizability and multiple composability. On the one hand, it is quite possible that microphysically different networks (across different brains, for example, or even in radically different kind of physical materials) are capable of realizing a set of vehicles that are subject to the same set of distance relations. On the other, we see how it is possible for the same material mechanism to generate a vast range of structurally distinct vehicles.

This last fact about neural activation patterns is critical when thinking about one of the major puzzles about consciousness. The neural correlates of very different conscious experiences, at least at one level of description, can look very much the same—at the base level each of these experiences is correlated with a pattern of neural activity. The puzzle is how this one substrate is capable of generating such a diverse range of conscious experiences. But the framework we have just been exploring offers us a way of resolving this puzzle. This one neural substrate is capable of generating very different kinds of conscious experiences because it is capable of generating a vast range of structurally distinct activation patterns.[15]

Applying a structuralist ontology to consciousness has a number of appealing features. It enables us to unify the science of consciousness with the other special sciences. On this story, conscious experiences are emergent structures just like molecules, airfoils, and cells. It also allows us to escape the pitfalls of both microphysicalism and functionalism. The existence of emergent structures is consistent with both multiple realizability and multiple composability. And because such structures are intrinsic to their realizing vehicles, they make sense of the causal relations into which these vehicles enter. This middle-ground alternative to both microphysicalism and functionalism would seem to get things just right.

References

Bechtel, William (2006). *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. New York: Cambridge University Press.

Bechtel, William (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Lawrence Erlbaum.

Bechtel, William & Adele Abrahamsen (2002). *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks, 2nd Edition*. Oxford: Basil Blackwell.

Brown, Michael E. & Gerry E. Swallowe (1981). The thermal decomposition of the Silver (I) and Mercury (II) salts of 5-nitrotetrazole and of Mercury (II) fulminate. *Thermochimica Acta*, 49, 333-49.

Bunge, Mario (1997). Mechanism and explanation. *Philosophy of the Social Sciences*, 27(4), 410-65.

Chalmers, David (1996). *The Conscious Mind*. New York: Oxford University Press.

Churchland, Paul M. (1995). The Engine of Reason, the Seat of the Soul. Cambridge, MA: MIT Press.

Craver, Carl F. (2001). Roles, Functions, Mechanisms, and Hierarchy. *Philosophy of Science,* 68, 31-55.

---

[15] See O'Brien & Opie 1999 for further discussion of this point.

Craver, Carl F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. New York: Oxford University Press.

Cummins, Robert (1975). Functional Analysis. *The Journal of Philosophy,* 72(20), 741-65.

Faust, W. L. (1989). Explosive molecular ionic crystals. *Science,* 245(4913), 37+.

Fodor, Jerry A. (1974). Special sciences. *Synthese,* 28, 97-111.

Fries, Pascal, Danko Nikolić & Wolf Singer (2007). The gamma cycle. *Trends in Neurosciences*, 30(7), 309-16.

Glennan, Stuart (2002) Rethinking Mechanistic Explanation. *Philosophy of Science*, 69, S342-53.

Hempel, Carl (1966) *Philosophy of Natural Science*. Prentice Hall.

Köhler, Wolfgang (1920). Die physischen Gestalten in Ruhe und im stationären Zustand. Braunschweig: Vieweg und Sohn.

Kurzer, Frederick (2000). Fulminic acid in the history of organic chemistry. *Journal of Chemical Education*, 77, 851-7.

Lamme, Victor A. F. (2004). Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural Networks,* 17, 861–872.

Lamme, Victor A. F. (2000). Neural Mechanisms of Visual Awareness: A Linking Proposition. *Brain and Mind,* 1, 385–406.

Llinas, Rodolfo, U. Ribary, D. Contreras & C. Pedroarena (1998).The neuronal basis for consciousness. *Phil.Trans. R. Soc. Lond., B*, 353, 1841-49.

Machamer, Peter, Lindley Darden, & Carl Craver (2000). Thinking about Mechanisms. *Philosophy of Science*, 67, 1–25.

McGinn, Colin (1990). *The Problem of Consciousness*. Oxford: Blackwell.

Miller, Steven M. (2007). On the correlation/constitution distinction problem (and other hard problems) in the scientific study of consciousness. *Acta Neuropsychiatrica*, 19, 159–76.

Nagel, Thomas (1974). What Is it Like to Be a Bat? *Philosophical Review*, 83, 435-50.

Newman, James (1995). Thalamic contributions to attention and consciousness. *Consciousness and Cognition*, 4(2), 172-93.

O'Brien, Gerard. & Jon Opie (1998) The Disunity of Consciousness. *Australasian Journal of Philosophy*, 76, 378-95.

O'Brien, Gerard. & Jon Opie (1999) A Connectionist Theory of Phenomenal Experience. *Behavioral and Brain Sciences*, 22, 127-48.

O'Brien, Gerard & Jon Opie (2000) Disunity defended: A reply to Bayne. *Australasian Journal of Philosophy*, 78, 255-63.

Oppenheim, Paul & Hilary Putnam (1958) Unity of Science as a Working Hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (eds), *Minnesota Studies in the Philosophy of Science*, 2. Minneapolis: University of Minneapolis Press.

Priestley, Joseph (1775) *History and Present Status of Electricity, 1, 3rd ed*. London: C. Bathurst & T. Lowndes et al.

Revonsuo, Antti (2006). *Inner Presence: Consciousness as a Biological Phenomenon*. Cambridge, MA: MIT Press.

Revonsuo, Antti (2010). *Consciousness: The Science of Subjectivity*. Psychology Press.

Robertson, Teresa (2008). Essential vs. Accidental Properties. *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, E. N. Zalta, ed., URL = <http://plato.stanford.edu/archives/fall2008/entries/essential-accidental/>.

Rumelhart, David, James McClelland & the PDP Research Group (1987) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations*. Cambridge, MA: MIT Press.

Salmon, Wesley (1978). Why ask "Why?"? *Proceedings and Addresses of the American Philosophical Association*, 51, 683-705.

Salmon, Wesley (1989). Four Decades of Scientific Explanation. In P. Kitcher & W. Salmon (eds), *Minnesota Studies in the Philosophy of Science, 13*. Minneapolis: University of Minneapolis Press.

Singer, Wolf (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24, 49-65.

Singer, Wolf (2007). Binding by synchrony. *Scholarpedia*, 2(12),1657.

Smart, John (1959). Sensations and Brain Processes. *Philosophical Review*, 68, 141-56.

Tononi, Giulio (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5:42 1-22

Tononi, Giulio (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215, 216-42.

Tononi, Giulio & Christof Koch (2008). The neural correlates of consciousness: An update. *Annals of the New York Academy of Sciences*, 1124(1), 239–61.

Varela, Francisco, J-P. Lachaux, E. Rodriguez & J. Martinerie (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2, 229-39.

Waldmann, Rainer, Guy Chamigny, Frédéric Bassilana, Catherine Heurteaux & Michel Lazdunski (1997). A proton-gated cation channel involved in acid-sensing. *Nature*, 386(6621), 173-7.

Woodward, James (1989). The causal mechanical model of explanation. In P. Kitcher & W. Salmon (eds), *Minnesota Studies in the Philosophy of Science, 13*. Minneapolis: Uiversity of Minneapolis Press.

Zeki, Semir & Andreas Bartels (1999). Towards a theory of visual consciousness. *Consciousness and Cognition*, 8(2), 225-59.

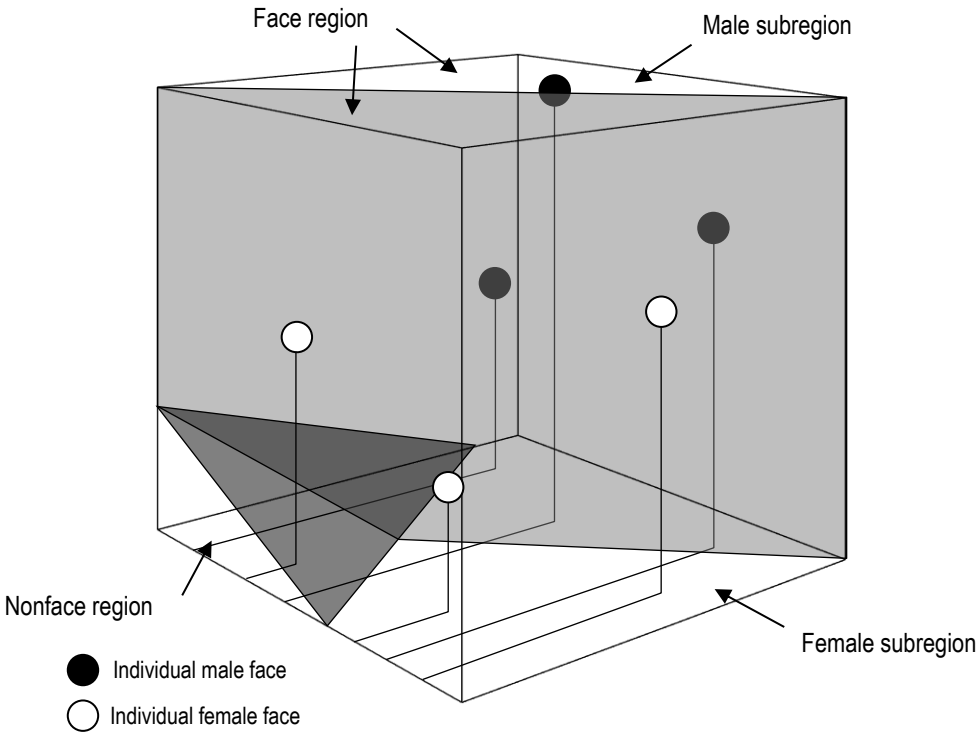

Figure 1.The hierarchy of learned partitions across the hidden unit activation space of Cottrell's face recognition network (after Churchland 1995, p.49).