History and Philosophy of the Life Sciences – draft version
(final paper: HPLS (2018) 40: 40. https://doi.org/10.1007/s40656-018-0205-2)

The search of "canonical" explanations for the cerebral cortex

Alessio Plebe

Abstract This paper addresses a fundamental line of research in neuroscience: the identification of a putative neural processing core of the cerebral cortex, often claimed to be "canonical". This "canonical" core would be shared by the entire cortex, and would explain why it is so powerful and diversified in tasks and functions, yet so uniform in architecture. The purpose of this paper is to analyze the search for canonical explanations over the past 40 years, discussing the theoretical frameworks informing this research. It will highlight a bias that, in my opinion, has limited the success of this research project, that of overlooking the dimension of cortical development. The earliest explanation of the cerebral cortex as canonical was attempted by David Marr, deriving putative cortical circuits from general mathematical laws, loosely following a deductive-nomological account. Although Marr's theory turned out to be incorrect, one of its merits was to have put the issue of cortical circuit development at the top of his agenda. This aspect has been largely neglected in much of the research on canonical models that has followed. Models proposed in the 80s were conceived as mechanistic, with the identification of a small number of components interacting in a basic circuit, and the definition of the functions of each component. More recent models have been presented as idealized canonical computations, distinct from mechanistic explanations, due to the lack of identifiable cortical components. Currently, the entire enterprise of coming up with a single canonical explanation has been criticized as being misguided, and the premise of the uniformity of the cortex has been strongly challenged. This debate is analyzed here. The legacy of the canonical circuit concept is reflected in both positive and negative ways in recent large-scale brain projects, such as the Human Brain Project. One positive aspect is that these projects might achieve the aim of producing detailed simulations of cortical electrical activity, a negative one regards whether they will be able to find ways of simulating how circuits actually develop.

A. Plebe
Department of Cognitive Science, v. Concezione 8 Messina Italy
E-mail: alessio.plebe@unime.it

Keywords cerebral cortex · canonical circuit · canonical computation · mechanistic explanation · cortical column · cortical development

1 Introduction

Neuroscience is invested with the highest cultural expectations. It is expected to answer questions such as: What do we think with? What is consciousness? Questions such as these have set the agenda for deciphering a small layer of the brain, the cerebral cortex. This thin outermost sheet of neural tissue that covers the brain is thought to be where higher cognition takes place in humans and the site of consciousness. (Miller et al, 2002a; Farhat, 2007; Fuster, 2008; Nieder, 2009; Noack, 2012).

Although it is not easy to define in a sentence or two, what it is exactly that distinguishes mammal intelligence from that of other species, the cortex is certainly considered to be the crowning achievement of brain evolution. The quest for an understanding of its function is among the most prominent and yet unresolved issues in neuroscience. This quest has developed into several distinct research domains, such as area localization (Brodmann, 1909), identification of cellular components (von Economo and Koskinas, 1925), and has provided evolutionary (Nauta and Karten, 1970) as well as developmental (Rakic, 1971) accounts.

The aim of this paper is to explore a concept, that of a "canonical" core underlying the processing power of the cerebral cortex, first articulated in the 70s, and turned into a more determinate research objective two decades later. I will trace the history of this line of research, and characterize its methods and achievements under a philosophy of science perspective. In particular, I will evaluate how canonical proposals fit in the framework of mechanistic explanations, the currently most accepted framework in biology (Glennan, 1996; Craver and Darden, 2013; Braillard and Malaterre, 2015b), and specifically in neuroscience (Craver, 2007).

There is an initial aspect that emerges from the research on the canonical core of the cortex, one that focuses on computational properties of the cerebral cortex, rather than, for example, its anatomical or cytological properties. As we will see in historical perspective, this domain of research has been enriched by the theoretical stance that emerged in neuroscience in the 60s and 70s, and was particularly receptive of the influence coming from the technical advances in electronics of the time. In turn, research on the canonical core of the cortex contributed substantially to the establishment of the theoretical and computational perspective in neuroscience, which is certainly not unique to this line of research (Churchland and Sejnowski, 1994; Dayan and Abbott, 2001).

There is, arguably, a more distinctive feature in the explanatory aim of this research, as gleaned from the scrutiny of the many proposals of a canonical core over the past four decades. It hinges on two apparently clashing observations, the remarkable uniformity in structure of the cortex, on one hand, and the bewilderingly variety of functions it supports, on the other. Let U/V from here

onward refer to this paradox. A way out of U/V is the idea that a "canonical" computational strategy has evolved in the cortex, one so effective that it can solve many seemingly disparate types of information processing tasks.

In section 2 I will address the variety of functions performed by the cortex, one that is ample enough to fully justify the "V" in U/V. The issue of uniformity is more controversial, and will be discussed in section 3, where the tentative conclusion is that the cortex is indeed sufficiently uniform in its nature to justify the search for a canonical circuit. In reviewing the history of canonical models, I will introduce the reader to David Marr (1970), whose theory is described in section 4. Subsequently, canonical solutions construed as *circuits* are described and evaluated in section 5. I will conclude, in section 6, with a review of a number of proposals in terms of idealized computations.

All throughout the history of canonical research outlined here, I acknowledge the important merits it has had in progressing our understanding of the cortex, and its success in identifying several regularities in its workings. I will argue, however, that all canonical proposals have failed to explain U/V. According to my analysis, the reason for this failure is that these explanations have omitted the developmental dimension of cortical circuits. Ironically, this was well hinted at in the early work on canonical models done by Marr, but since then completely ignored. The focus of all subsequent research has almost entirely been on mature circuits and functions, neglecting the enormous capacity of the cortex to mold its computational functions in response to input patterns. There are objective difficulties in evolving basic cortical circuits through "canonical" development rules, but the need for new efforts in this direction is now acknowledged by many of the main players in the field of canonical research, and can represent a promising future path to take in resolving many of the U/V issues of the cortex.

2 Functions, mechanisms, and computations

In the variety of canonical proposals available U/V has not always been the main explanatory target. Deflationary accounts exist where "canonical" only means that a certain circuitry sketch of a particular computation, recurs often in the cortex (and maybe even in other parts of the brain). It is possible that the observed canonical feature by itself does not play an explanatory role. However, for all canonical proposals that construe a model of circuits or computation as a basic feature of the cortex, there are reasons to expect U/V to be an *explanandum*. The most fundamental reason is that since uniformity in structure and a high variety of functions are the remarkable and distinctive features of the cortex, a canonical solution that aims to capture the essential way the cortex works is also committed to trying to explain U/V.

The "U" side of U/V will be addressed in the next section. In dealing with the V side, the first concern is that in asserting that the cortex is able to perform a wide variety of functions requires being precise about what a "function" might be for the cortex. This question has been extensively dealt with by

philosophers of various disciplines that include the cortex among their objects of investigation, such as biology, neuroscience, psychology, and philosophy of mind. Most of the different positions have their origin in two fundamental perspectives. The etiological theory of functions proposed by Wright (1976) holds a realistic concept of functions, as a task nature has built some system to perform. Against this realistic claim, Cummins (1975) defended the idea of functions as capacities of the components of a system, that have a causal role in the system, with respect to an explanatory strategy. In the last 40 years, both theories have stimulated a wealth of expansions and new directions, the most relevant for our discussion being the synthesis of Cummins' causal role analysis with mechanistic explanation (Machamer et al, 2000; Craver, 2001), aimed at identifying entities and activities ("capacities" for Cummins) in a system, organized in the production of regular changes in the system.

Etiological accounts find easy acceptance in biology in general, because of the essential relation of biological components with natural selection, with refinements and variations ever since Wright (Garson, 2016). Physiology is somewhat different. The causal role view of functions is in better agreement with the methodology of the discipline, focused on the analysis of a living system, independently from its past history, and with clinical interest: the understanding how functions can fail and how those failures might be predicted and controlled (Roux, 2014). For similar reasons, causal role theory is the prevailing account in neuroscience as well, especially due to the close relation with mechanistic explanation (Craver, 2007), and with the additional advantage of offering a notion of function close to the computational account, which will be discussed later on. On the other hand, etiological theories revived a new line of research in philosophy of mind, with the teleosemantic theories of mental content (Millikan, 1989; Papineau, 1993). In the end, the cerebral cortex may stand at the crossroad between both views of "function". On the one hand, because, generally speaking the cortex is a biological system, it does what natural selection pressures require of it. On the other hand, the cortex has a major casual role in most – if not all – of the behaviours of the animal.

In fact, an endless conundrum in cognitive neuroscience concerns the localization of "functions", in the sense of cognitive capacities (often defined as etiological functions), in the cortex. Young et al (2000) argued that the conventional imputing of cognitive function to particular cortical areas derived from the effects of lesions is unreliable, if not supplied with wider connectivity analysis. The traditional view of the cortex as being organized in modules that bear correspondence with cognitive functions has been challenged by many Prinz (2006); Bergeron (2007); Burnston (2016b). Several have challenged the association between cortical regions and "functions" as cognitive capacities, suggesting alternative association for cortical regions such as "intrinsic functions" Rathkopf (2013) or "difference-makers" in relation to cognitive functions Klein (2017).

However, for the purposes of our argument, it is easy to verify the cortex's involvement in a multitude of functions under all relevant accounts of the term "function". When we adopt cognitive capacities as "functions", it

is difficult to conceive a capacity whose performance would not involve the cortex in a substantial way. This in fact can be said for vision (Chalupa and Werner, 2003); somatic perception (Nelson, 2002a);language (in humans) (Pulvermüller, 2002); planning and decision making(Fuster, 2008); consciousness (Dehaene, 2014) and moral cognition(Verplaetse et al, 2009). Of course, all of these functions involve subcortical nuclei as well, but the dominant role the cortex plays in all of these capacities, at least in mammals, is widely accepted.

When adopting the physiological view of "function", a great help in characterizing brain functions comes from computational and theoretical neuroscience (Dayan and Abbott, 2001; Rathkopf, 2013; Burnston, 2016a). Functions, in this sense, regard the mapping of input signals onto output signals and the (mathematical) transfer function involved in this process. Well established examples in the cortex include the realization of Gabor transfer functions by receptive fields in the primary visual cortex (Jones and Palmer, 1987); bimodal response distributions with respect to orientations in area V2 (Anzai et al, 2007; Plebe, 2012); spectrotemporal signal transformation functions operated by receptive fields in primary auditory cortex (Miller et al, 2002b); and encoding of familiarity as skewness in hippocampal place fields (Mehta et al, 2000).

The computational account deserves further theoretical clarification. All proponents of canonical explanations for the cortex adhere to the view that computation is not just a useful tool for modeling and simulating cortical functions, but believe that the cortex itself computes, using forms of computation much different to those run on man-made computers. Similarly to what we have seen for "function", computation, for what concerns the cortex, is contended between two domains. On one side in cognitive science *Computational Theory of Mind* (Fodor, 1975; Pylyshyn, 1981) has advocated for cognitive computation the same foundation of "computation" found in theoretical computer science (Turing, 1936; Church, 1941), on the other side in neuroscience, roads far from the Turing machine model have been taken,that are instead tied to the brain's physiology (Wilson and Bower, 1989; Hines and Carnevale, 1997; Bower, 2005). In the last decade, significant progress has been achieved in theoretical accounts of physical computation that reconcile both man-made digital computers and brains (Copeland et al, 2013; Miłkowski, 2013; Fresco, 2014; Piccinini, 2015).

Moreover, since both computational and mechanistic explanations seem proper to neuroscience (even more so for the cortex), efforts have been directed at analyzing conditions for the two approaches to meet. Recently, several criteria have been proposed for ascertaining, which computational model may provide the proper mechanistic explanation of the modeled system (Piccinini, 2006, 2007). More specifically, Kaplan (2011); Kaplan and Craver (2011) propose the 3M (*model-mechanism-mapping*) constraint on computational models that are claimed to have (mechanistic) explanatory power:

> A model of a target phenomenon explains that phenomenon to the extent that (a) the variables in the model correspond to identifiable components, activities, and or-

> ganizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these (perhaps mathematical) variables in the model correspond to causal relations among the components of the target mechanism.

Of course, this constraint does not work as a strictly true or false test, because complete mechanistic models of neural behavior are unfeasible to construct in practice. Even so, the definition can also be applied to incomplete models, where details are omitted either for reasons of computational tractability or because these details are still unknown. It is thus, mostly a guiding principle: computational models whose design does not specify a structural correspondence between model components and external system components are merely predictive models, while those designed with such a correspondence can be evaluated as explanatory models. This criterion is not universally agreed upon, and as we will see next, there are cases of quite abstract canonical computations, for which a specific explanatory power is claimed.

3 The uniformity of the cortex

This section considers claims for the uniformity of cortical structure, as expressed in the premise of arguments, quoted in the Introduction, for the existence of a canonical basic kernel for the cortex. This claim is intrinsically problematic because it asserts a quantitative property without any explicit metric by which the claim can be evaluated. Of course the cortex is not uniform down to the molecular level like a metal plate; cortical regions clearly differ in their details, both between different areas of the same individual and between homologous regions in different individuals of the same species. That said, the structure of the cortex *is* remarkably uniform in comparison to other areas of the nervous system, such as the diverse nuclei in the brainstem. In the sections below, we will evaluate the uniformity of specific spatial dimensions and scales separately, comparing radial (vertical) variation with intralaminar (horizontal) variation.

3.1 The layered structure

The most well-investigated kind of uniformity is the regular repetition of the radial profile of the cortex. As first observed by Berlin (1858), cortical tissue can be grouped into distinct layers parallel to the cortical surface. Until the beginning of the last century the exact number of layers was controversial (Smith, 1992). Theodor Meynert (1869), using Berlin's staining methods, found that the most common lamination consists of five layers, although the number can vary up to eight depending on the area of the cortex. Five was also the number found by the Russian anatomist Vladimir Betz (1874) using carmine staining. In reviewing the figures claimed by his contemporaries, Ramón y Cajal (1891) leaned towards four.

The delineation of six layers became the orthodox convention when Brodmann (1903, 1909) and Vogt and Vogt (1903) adopted a new cytoarchitectonic method, a meticulous analysis of Nissl preparations stained for nerve cells. Vogt and Vogt (1919) used also myeloarchitectonics, the study of the course of the horizontal, intracortical myelinated fibers stained with the Weigert method, and Braak (1974) introduced pigmentoarchitectonics, based on a selective staining technique for lipofuscin pigments. All these methods converged in a blueprint for a six-layered structure unfolding all over the cortex, with several minor variations. The main difference is in the fourth layer, called *granular* because of its population of small and packed spiny stellate cells. The input to spiny stellate neurons is fed primarily by thalamic fibers, and thus layer IV is prominent in all sensorial areas, collectively called *granular* cortex, while less so in motor areas, called *agranular*. At a finer level, cortical areas are usually classified according to several more subtle differences, for example areas with large pyramidal cells in layer III, like the inferofrontal, are called *magnopyramidal*; areas with giant pyramidal cells in layer V, like the intermediate somatomotor cortex, are called *gigantoganglionic*; and areas like the retrosplenial lateral region, where small cells prevail are called *parvocellular*. Despite all these local variations, the basic layered structure is preserved.

3.2 Invariance of neuroanatomic statistics

In a widely cited first attempt to assess the uniformity of the cortex on a quantitative basis, Rockel et al (1974, 1980) counted the number of cells through the entire thickness of the cortex in most of the major cortical areas in monkeys, humans, and several other mammals. They found this count to be surprisingly constant across both areas and species, with about 110 neurons in cortical sections of $30\mu m$ in diameter. In each species, the exception is the primary visual cortex, with a count of about 270 neurons. Their observations have been the subject of fierce debate for over 30 years, with doubts raised concerning whether their experimental methods were technically flawed (Rakic, 2008), and other studies reporting twofold or even threefold variation in neural density across cortical regions (Herculano-Houzel et al, 2008). Carlo and Stevens (2013) recently replicated the direct count performed by Rockel et al. using modern stereological methods, confirming the same uniformity of count for the same species and cortical areas. The absolute number of neurons they found is 14% less than the number reported in the 1980 study, but the constancy of counts across the cortex was confirmed. Karbowski (2014) reviewed and compared a number of neuroanatomical features of the cortex, in mammals ranging from humans to dolphins, and again found a remarkable invariance across species and across regions. In humans the mean length of postsynaptic density, the thick part of the postsynaptic membrane hosting neurotransmitter receptors, is of $0.38\mu m$ with a standard deviation of only $0.04\mu m$, and the synaptic density has a mean of $5 \times 10^{11}\text{cm}^{-3}$ with standard deviation 0.3.

The ratio of excitatory to inhibitory synapses is highly invariant even across species, with an average of 0.83 and a standard deviation of 0.03.

3.3 The column, a canonical competitor

In addition to the qualitative and statistical uniformity of the radial organization, a third kind of possible uniformity relates to the surface organization of the cortex. Evidence in many cortical areas suggests periodical replication of a local cortical circuitry parallel to the surface, has given rise to the so called *columnar* organization of the cortex. The word "column" (*Säule*) was first used by von Economo and Koskinas (1925), and the idea that tiny cylinders of cells oriented radially might be a common unit in the cortex was suggested by Lorente de Nó (1938). This assumption was first demonstrated by Mountcastle (1957) in the somatic sensory cortex, with functional evidence for vertical cylinders, no more than 500μm wide, of neurons all responding to stimulation of cutaneous receptors located at a particular site. Further evidence came a few years later with the discovery of columnar organization in the primary visual cortex (Hubel and Wiesel, 1959). According to Haueis (2016) the cortical column idea of Mountcastle offered a fruitful *conceptual outlook* to later investigations such as those of Hubel and Wiesel, but "this outlook only briefly took a form that one could call a 'theory' of the cerebral cortex, before new experimental techniques started to diversify column research."

In fact, the idea that the entire cortex was composed of a patchwork of alternating columns of neurons is difficult to reconcile with the specific columnar organizations that have been observed. For instance, in the primary visual areas, ocular dominance columns are about 500μm in size, yet orientation columns in the same region are 100μm (Hubel and Wiesel, 1963), much smaller than in the somatic sensory cortex. Another crucial concept was introduced by Rakic (1995), the "ontogenetic column", a vertical stack of cells, divided by glial septa, generated during the embryonic migration of neurons into the cortical plate. This column was revealed to be smaller than Mountcastle's columns by an order of magnitude. This diversity in column sizes has been addressed by extending the nomenclature, calling a "minicolumn" the adult version of Rakic's ontogenetic column (Mountcastle, 1997), renaming "macrocolumn" the original, simple, functional column (Buxhoeveden and Casanova, 2002), and adding "hypercolumn" to refer to a complete rotation in primary visual orientation and eye preference domains (Hubel and Wiesel, 1974). This rather baffling proliferation of putative cortical surface organization led Horton and Adams (2005) to believe that "Although the column is an attractive concept, it has failed as a unifying principle for understanding cortical function."

The discussion on columns is currently open and still highly debated. A conservative view is that columnar organization is a fundamental feature of the cortex, even if not homogeneous and common to all areas (Kaas, 2012). The problem of the diversity of scales at which modular circuits reiterate in the cortex is one of the main issues raised by Horton and Adams. According

to Rothschild and Mizrahi (2015) the functional uniformity along the radial direction, the fundamental defining feature of columns, is maintained at a coarse topography of cortical maps, but map architecture at microscale becomes heterogeneous along cortical depth. There are attempts to save the idea of universality in column organization, by moving down in scale, as for example in the collection edited by Casanova and Opris (2015), who state in their introduction the following (p.4):

> It is the aim of this book to bring together observations in regards to cortical modules, from different research perspectives, that can generalize to other fields of science. We will emphasize the role of the smallest module capable of information processing: the minicolumn. Neurons are not reiterative elements of the brain. [...] One gains insight into the workings of the brain only by looking outside of the neuron and into minicolumns.

At the end of a critical review that shares several of the arguments of Horton and Adams, with more details on the development of aspects of the columns Molnár (2013) complains that (p.125-126) "The term column is still used because it is a captivating concept. For the time being, there is no easy alternative to column. It is necessary to establish more specific terminology that will allow specific reference to particular entities." An even more concessive recommendation is expressed by Rockland (2011), specifically addressing the classical macrocolumns: "As a term, column is imperfect. [...] Unfortunately, there is no easy alternative to column, and no more specific terminology. [...] For now, best may be to continue using the term." From a brief bibliometric verification it seems that the recommendation made by Rockland is still being followed, as shown in Fig. 1 there is a steady usage of "column" in hundreds of papers published per year on *Cerebral Cortex*, while the alternatives "minicolumn" or "microcolumn" appear in just a few papers per year.

It is useful to highlight here that the columnar hypothesis is more than additional support for the uniformity of the cortex, it already embeds a suggested basic circuit, that in some way competes with the "canonical circuit" idea. Therefore, the difficulties in reconciling the organization of the entire cortex under the columnar concept, rather that infringing the claim of the uniformity of the cortex, has helped the canonical concept in this competition. Maçarico da Costa and Martin (2010) suggest that "In moving away from this rather static image of the functional architecture [the columnar model] to the idea of repeated canonical circuits, it is not a great leap of the imagination to suppose that all of cortex carries a similar computation on its inputs, whether it be for perception, or more complex cognitive judgments". The challenge then is to explain how a single canonical circuit, if it exists, could nonetheless lead to such a variety of different columnar organizations in the various cortical regions.

The competition between the column and the canonical circuit can be regarded as contending supremacy between anatomy and physiology, or – in other words – structure and function, with Maçarico da Costa and Martin (2010) rallying for the latter. However, one should always remember that

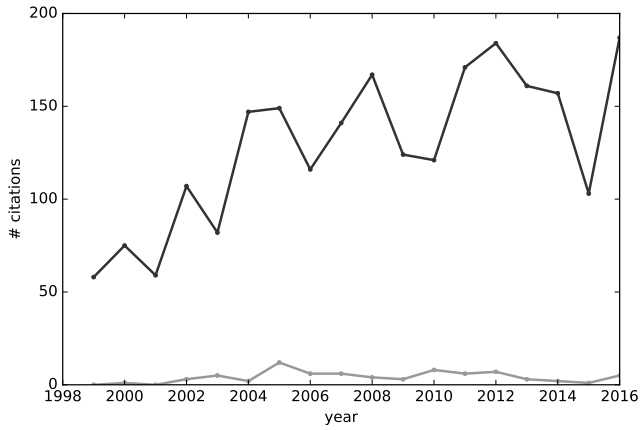

Fig. 1 Number of papers in the journal *Cerebral Cortex* using the term "column" (dark line), and the terms "minicolumn" or "microcolumn" (light line).

anatomy and physiology are partitions of science, not nature. These two separate disciplines emerged for historical reasons, due to the development of different sets of methods, and subsequent conceptual specialization, but the object of investigation is the same indivisible biological entity.

In sum, regardless of the diversity at the columnar level, there is still sufficient uniformity at the radial/laminar level and at the level of statistical regularities to warrant searching for a canonical circuit. In characterizing the clashing contrasts between the uniformity of the cortex and the wide range of different functions, Harris and Shepherd (2015) adopt the term "serial homology", which is used in physiology for the similarity in the organization of different structures within a single organism. For example, bones are very similar in their cellular structure, but differ depending on size and mechanical properties, such as in a femur or a phalange. Unlike bones, the cortex differentiates in computational rather than mechanical functions, achieved – it is argued – by a common canonical circuity, rather than compact osseous tissue.

4 The early attempt by David Marr

David Marr (1970) proposed the first canonical model in a long (and somewhat difficult to read) paper that is one of his least known works. It is part of a trilogy that he wrote as a doctoral student in theoretical neuroscience, under the supervision of Giles Brindley at Trinity College, Cambridge. The question he chose to address in his work was both a general and ambitious one, a theoretical speculation on how the brain works, meticulously compared with

the anatomical data available at the time. The adventuring of Marr into theoretical neuroscience, decades before becoming established, certainly resulted from his personal genius, and was favored by his mathematical background, and his being involved in the experimental and applicative neuroscientific activity in the Physiological Laboratory in Cambridge led by Brindley, a pioneer in neural prostheses (Brindley and Lewin, 1968). The theoretical exploration of the brain with mathematical tools was not on Brindley's agenda, but was certainly within his insatiable scientific curiosity. He occasionally attended the Ratio Club as a guest, a British cybernetic dining club whose members included scientists such as Ross Ashby and Alan Turing (Holland and Husbands, 2011).

In his essay Marr did not use the term "canonical", instead calling his proposal the "fundamental neural model" of the cortex. The meaning, however, is clearly the same, and the motivation for its existence was precisely the U/V problem:

> The mammalian cerebral neocortex can learn to perform a wide variety of tasks, yet its structure is strikingly uniform. It is natural to wonder whether this uniformity reflects the use of rather few underlying methods of organizing information
>
> (Marr, 1970, p.163)

The work of Marr is organized in two distinct parts: the first, covering sections 0-3, proceeds in pure abstract mathematical terms, the second, sections 4-7, attempts to derive correspondences between entities of the mathematical model and cells in the cortex. The first part loosely follows a deductive-nomological argument (Hempel, 1965), with the definition of the phenomenon, its characterization by abstract laws, and the derivation of mathematical models, although most derivations are not formalized. The phenomenon is the ability of a sentient agent to classify an event E_i of the world as Ω_j, where each class Ω_j is a type of event of relevance for the organism. He used as an example of a class the set of events that represent various types of poodle. The classification is based on a set of detectable features $a_{i,k}$, probabilistically associated with the event instance E_i. Two events that represent different poodles E_1 and E_2 should have more features $a_{1,k}$ and $a_{2,k}$ in common, than the number of features shared by E_1, and an event E_3 that represents something different from a poodle.

The basic law behind this problem is the so-called Fundamental Hypothesis, spelled out as (p.182-183):

> Where instances of a particular collection of intrinsic properties (i.e. properties already diagnosed from sensory information) tend to be grouped such that if some are present, most are, then other useful properties are likely to exist, which generalize over such instances. Further, properties often are grouped in this way.

From this basic law several theorems are informally derived, such as the "diagnosis theorem", that relates the conditional probability $P(\Omega_j|E_i)$ of the event instance E_i to be of type Ω_j with the functions $c_k(E_i)$, called "evidence

function", returning 1 if the feature $a_{i,k}$ is present, 0 otherwise. The "interpretation theorem" deals with cases in which, only a subset of features $a_{i,k}$ become available during an event instance.

After a detailed formulation of this general theory of how mathematical models might explain classification of sensory signals, Marr engaged in an audacious attempt to adapt the working of the cortex to this theory (p.162):

> The structure of the cerebral neocortex is reviewed in the light of the model, which the theory establishes. It is found that many elements in the cortex have a natural identification with elements in the model.

Marr used the term "codon", borrowed from molecular genetics, to address elements encoding subsets of features, and speculated that Martinotti (1890) cells may be plastic codon cells, and spiny stellate cells in layer IV could be codons with predefined output classes of events. He provided several diagrams of template circuits, based on Martinotti, spiny stellate, and pyramidal cells, that fulfill his theory. Marr completed the mapping with a series of schemes of the interconnections between the neural elements (Figures 6, 7, and 8 in his original paper), where their causal connections match the mathematical relations in his model. For example, there were pyramidal cells whose output matches with the function $P(\Omega_j|c_i)$, with c_i the activation of an excitatory codon cell synapsing to the pyramidal cell.

Needless to say, this ambitious attempt to derive an organization at the level of neural circuits from abstract mathematical laws was hopeless. The overall strategy of Marr was synthetic, developing an explanation from a conjecture of how the behavior of the isocortex might follow mathematical laws, and then identifying components responsible for the several subtasks. This strategy was exposed to the risk of producing spurious explanations. One source of error was what Bechtel (1982) called "vacuous functional analysis", the way a researcher thinks a system might produce its activity, whose steps do not describe processes actually performed by the system. As diagnosed by Bechtel (p.560) "this type of error typically arises when a researcher focuses on evidence on how the system actually behaves and tries to construct a model of how that behavior could be brought about." Marr was aware of this risk (p.196):

> The central difficulty with producing neural models for a specific function is that there are many ways of doing the same thing [...] so rigorous derivation of the basic neural models cannot proceed very far. This does not, however, commit the discussion to unredeemed vagueness. The injection at strategic points of a little common sense allows enough precision in the models

Given the complexity of the cortex, and the frugal amount of details available at the beginning of the 70s, hopes of hitting the bull's eye of its canonical core by way of a synthetic strategy were scarce even for the "common sense " of a brilliant scientist like David Marr. In the period between 1971 and 1972, he, himself, recognized the risks run by abstract general theories of the brain, that lacked an understanding of specific neural mechanisms. The risk was that of

resulting glaringly incomplete and almost fruitless. His shift in view is clearly stated in a review of the proceedings of a summer school held in Trieste in 1973, *Physics and Mathematics of the Nervous System* Marr (1975, p.875):

> Many experimental biologists dismiss with contempt the approach of even very able theoreticians in developmental or neurophysiological problems. The outsider need look no further than this volume to understand why. [...] papers describe attempts to elucidate problems of biological information processing, but in one way or another they all make the same strategic error – engaging in the search for a general theory before and actually instead of tackling the particular problems at hand. [...] With problems of biological information processing there has been almost no experience, and one's intuition is at best untrustworthy. It may even be that biological information processing admits of no general theories except those so unspecific as to have only descriptive and not predictive powers.

In this period, Marr himself, was following the direction recommended in this review. When he moved, in 1973, to the Artificial Intelligence Laboratory at MIT, he abandoned the speculative theoretical road, shifting his efforts to the study of the visual system, favoring bottom-up work grounded in an understanding of mechanisms involved in specific tasks.

Nevertheless, the ambitious early attempt of Marr had several merits, among which was that of ushering in the birth of the search of what would later be called the "canonical" cortical circuit. David Willshaw (1972) specifically acknowledged a debt to Marr in his development of an artificial network, called "Inductive Net", better able to deal with novel inputs rather than simple associative networks. Later, he further extended the work carried out by Marr, including learning rules for the synaptic strength of the connections that were underspecified in Marr's model (Willshaw et al, 1997).

I argue that one of the key intuitions of Marr was lost during the subsequent history of the cortical canonical core. As noted by Cowan (1991), Marr's theory of the cortex "tackles a much more complicated problem, that of the *ab initio* formation and organization of networks capable of classifying and representing the world." Marr was clear on the issue: "The mammalian cerebral neocortex can *learn* to perform a wide variety of tasks". This crucial aspect has since been largely ignored in canonical models, that assume development to have already taken place.

5 Canonical microcircuits

After Marr, the first proposals of canonical cores of the cortex came in the form of "circuits", and included drawings as electrical circuits. The blending of electrical and electronic engineering with neurophysiology in the last century is recognized as a significant moment in the epistemology of neuroscience (Brazier, 1961; Newcomb, 1994; Borck, 2001; Rose and Abi-Rached, 2013). A paradigmatic case of the impact progress in the field of electrical engineering has had in the field of neuroscience is the cable equation (Pettersen

and Einevoll, 2009). William Thomson, Lord Kelvin of Largs, was involved as scientific adviser in a project of epic proportions, that of the transatlantic telegraph cable. He derived from his previous theories of heat conduction (Thomson Kelvin, 1842) an equation describing the electric potential as a function of time and of the linear dimension of the cable, ruled by static electric parameters of the cable (Thomson Kelvin, 1855). This theoretical advancement overcame the failures of the first submarine cable, gaining Thomson the title of 1st Baron Kelvin. Decades later, Jan Hoorweg (1898) was one of the first to realize that Kelvin's equation might do a much better job at describing electrical behavior in motor neuron axons, than the laws of du Bois-Reymond (1849), but the latter's authority caused the quick dismissal of Hoorweg's idea. Half a century later Wilfrid Rall (1957, 1969) accomplished the job proposed by Hoorweg, introducing in neuroscience what is still called today the "cable equation", a direct adaptation of Kelvin's equation to neural membrane potentials. The same basic equation is at the heart of the NEURON simulator (Hines and Carnevale, 1997), still adopted in the currently most advanced detailed simulators of electrical activity in assemblies of neurons (Markram et al, 2015). Those kind of "circuits" inherit exactly the abstractions assumed in electrical engineering, as networks of idealized quantized components of very few types (batteries, resistors, inductors, capacitors), connected by ideal perfect conductor lines, and node connections obeying the laws of Kirchhoff (1845). The canonical circuits we will describe shortly, inherit the main assumption of electrical circuits, they approximate the electromagnetic field as described by Maxwell into a finite set of attributes that does not depend on their position in physical space (Paynter and Beaman, 1991). Unlike the cable equation, the elements in canonical circuits are not standard electrical components, but "neurons", abstracted in a few classes.

5.1 Two canonical microcircuits

Gordon Shepherd obtained his PhD from the University of Oxford, working with Charles Phillips and Tom Powell on the responses of mitral cells to stimulation of the lateral olfactory tract in the rabbit (Phillips et al, 1963). This research attracted the attention of Rall, who at the time had just established the Mathematical Research Branch of the National Institute for Arthritis and Metabolic Diseases in Bethesda, a small early group in theoretical neuroscience (Rall, 2006). He envisioned the possibility of building a theoretical model reproducing the action potentials of the mitral cells recorded in the experiments done by Phillips, Powell and Shepherd, and had invited Shepherd for postdoc research. The Mathematical Research Branch had a computing facility, a Honeywell 800 machine, on which it was possible to implement a version of Rall's cable equation tuned to simulate the potentials of the mitral cell (Rall and Shepherd, 1968).

In the years that followed, Shepherd directed his research on the olfactory cortex towards the derivation of a core circuit, that he named the "summary

diagram of synaptic relations of neurons in prepyriform cortex" (Haberly and Shepherd, 1973, p.795). The advantage of pyriform cortex (which receives olfactory input), is the laminar structure in four layers, simpler than the rest of the cortex. The hippocampus shares much of the organization of the olfactory cortex, with a reduced number of layers, and similar patterns of connections between pyramidal cells and interneurons, and Shepherd (1979) derived a basic diagram of the hippocampus that was not too different from his earlier prepyriform diagram. Shepherd's main target was a core circuit for the general mammalian cortex, driven by much the same motivation as Marr. His view was that there should be a core organization that is characteristic of the cortex as a whole, and of the cortex only Shepherd (1974, p.258): "[...] certain modes of information processing are possible with a cortical type of organization that are not possible with other, noncortical, types."

Shepherd (1988) spoke of a "basic circuit" meant to be common to the entire cortex, without using the specific term "canonical", but again his approach fits the current definition of that term. A few years ago he confirmed the motivation of the canonical microcircuit as an explanation of the U/V problem:

> The neocortex is the brain structure most commonly believed to give us our unique cognitive abilities. Yet the cellular organization of the neocortex is broadly similar not only between species but also between cortical areas. This similarity has led to the idea of a common 'canonical microcircuit' employing a similar computational strategy to process multiple types of information.
>
> (Harris and Shepherd, 2015, p.170)

Compared to Marr, Shepherd aimed at a model that was both much simpler and more closely related to the physiology of the cortex. This circuit, shown in Fig. 2 A and D, has two inputs: one from other areas of the cortex (CORT) making excitatory synapses on dendrites of a superficial pyramidal neuron P1, and an afferent input (AFF), terminating on a spiny stellate cell SS and dendrites of a deep pyramidal neuron P2. This input, or course, is present in granular cortex only. There are feed-forward inhibitory connections through a superficial inhibitory neuron SI, and feedback connections through a basket cell BC.

Rodney Douglas and Kevan Martin were from Cape Town, where they did their university studies, Douglas up to his PhD in physiology in 1985, while Martin had moved earlier to Oxford to pursue his PhD at University College. In 1980 David Whitteridge, who had just retired from the prestigious Waynflete Chair of Physiology in Oxford, became interested in the techniques for marking neurons with the horseradish peroxidase enzyme, combined with intracellular recording (Brown et al, 1980). He convinced the Medical Research Council to fund a pilot project to experiment similar techniques in the attempt to relate structure and function in the visual cortex, and recruited a team of young researchers, whose most brilliant components were Douglas and Martin. Martin was especially quick in mastering horseradish peroxidase marking, producing in just a few years an impressive amount of investigations on the visual

cortex (Martin and Whitteridge, 1982, 1984a,b; Kisvárday et al, 1985), that favored the decision of the Medical Research Council to create the Anatomical Neuropharmacology Unit under Whitteridge's leadership. One of the most successful achievements of the new unit was the model proposed by Douglas et al (1989) as the "canonical microcircuit" of the cortex, thereby introducing the term "canonical". From then on, it became the standard way of addressing any sort of common basis for the cortex.

There is no doubt that for both Douglas and Martin as well, canonical microcircuits would be putative explanations for the U/V paradox of the cortex, as expressed in their following quotes:

> The apparent uniformity of the neocortex has given rise to the speculation that [...] is designed to perform the same basic operation, or 'computation' as it is now fashionable to call it. Yet, over 100 years of research in neurology and physiology has shown that the entire cortex is parceled into many (perhaps hundreds) functionally distinct areas. If all these areas have the same basic structural components and organize them in a similar way, the only thing that would distinguish auditory cortex from motor cortex is that it has different sources of input and sends its output to different targets. [...] The tempting notion is then that nature's laboratory has hit on a process that enables it to use the same machinery for very different ends. If this attractive view is correct, the $64000 question is then: what is the cortex doing with its inputs?
>
> (Martin, 1988, p.639–640)

> The uniformity of the mammalian neocortex has given rise to the proposition that there is a fundamental neuronal circuit repeated many times in each cortical area
>
> (Douglas et al, 1989, p.489)

The canonical circuit of Douglas and Martin is shown in Fig. 2 B and E. Compared to Shepherd's approach, the main difference is in the use of a minimal set of neuron-like units, disregarding the spiny stellate cells, whose effect is taken into account in the pyramidal-like cells P1 and P2. The only non-pyramidal neuron is a generic GABA-receptor inhibitory cell. The model was implemented using rate-encoding of the outputs of the three units, and their excitatory or inhibitory action was computed as a change in membrane potential after a transmission delay. The output of each unit was a thresholded hyperbolic function of the average membrane potential, after a constant time relaxation. The tuning of the model's parameters was derived by intracellular recordings in the cat visual cortex (area 17), in response to short electrical pulses (0.2–0.4 msec) above the lateral geniculate nucleus. The use of pulse stimulation was introduced, among others, by Phillips et al (1963), and offered the advantage of easier response analysis, as was well established in engineering standard system analysis. In providing data for the canonical circuit, pulse stimulation has the additional advantage of making the circuit general, independent of the many different natural stimuli, which in several cortical areas are unknown. In the area of neural recordings horseradish peroxidase was used to enable later morphological identification. The measured cortical responses

were also used as comparison with the output of the canonical model, and the results of both the model and of the experiments were found to be in agreement.

The connection between electronic engineering and neurophysiology tightened when Douglas and Martin met with Misha Mahowald. She was a pioneer in the field of neuromorphic engineering, where the stuff "circuits" were made of was silicon, but where the circuital diagram was derived from biological nervous systems, made into "circuits" thanks to the efforts of scholars like Rall, Shepherd, Douglas, and Martin. Mahowald did her doctoral research at the California Institute of Technology under the guidance of Carver Mead, who had just invented neuromorphic analog VLSI design (Mead, 1989), and as result she developed the *Silicon Retina*, a single silicon chip reproducing the first stages of retinal processing (Mahowald, 1994). Attracted by her results, in 1992 Douglas and Martin called Mahowald to collaborate with them at Oxford, moving soon from the retina to the cortex, with a neuromorphic system implementing cortical positive feedback (Douglas et al, 1994). Douglas and Martin saw analog VLSI design as a radical alternative to conventional computer simulations, and the best realization of their canonical microcircuit (Douglas and Martin, 1992, p.229-230):

> We have begun to realize that biological computation is very different from digital computation and that the neocortex may indeed be the source of yet undiscovered strategies of processing. [...] One avenue we are exploring is that of analogue very large scale integration (VLSI;Mead,1989). [...] The physics of the analogue silicon circuits is very similar to that of neurons. [...] The technology of analogue VLSI thus solves the major problem of simulating the behavior of large numbers of realistic model neurons. The advantages of using such a method for exploring the functions of the canonical microcircuits of neocortex are immense.

The Medical Research Council had too narrow a scope to host advanced electronic design in combination with neuroscience, and in 1995 Douglas, Martin and Mahowald moved to Zurich to establish the Institute für Neuroinformatick, jointly managed by the University of Zurich for the neuroscientific side, and ETH, the Swiss Federal Institute of Technology for the electronic engineering side. Sadly, the year after, Mahowald died unexpectedly, and the impact of her untimely demise was probably reflected in the temporary decline of interest in neuromorphic engineering at the Institute für Neuroinformatick, especially by Douglas and Martin. The two went back to working on the abstract canonical circuit, with further validations on other data, and minor revisions to the relative strengths of the connections (Douglas et al, 2004). The dominant excitation was now provided by intracortical connections between pyramidal neurons, so that even a relatively weak thalamic input could be greatly amplified. Even if inhibition was relatively weak, by modulating the recurrent excitation, it could still play an important role.

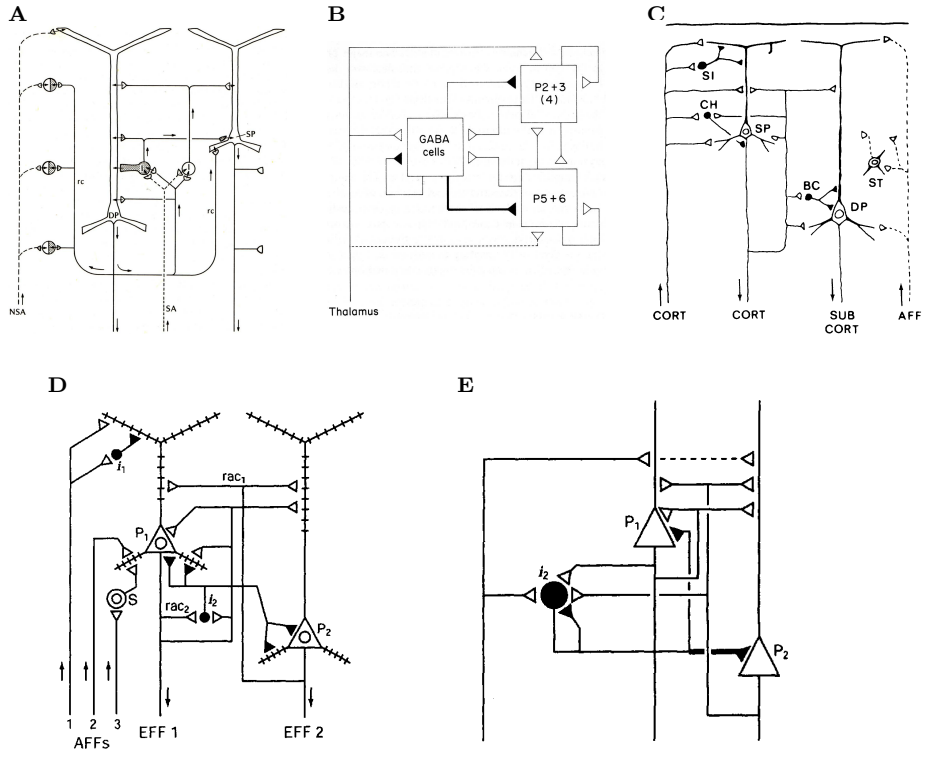

Fig. 2 Schematics of possible canonical circuits in the cortex. **A** is the circuit proposed by Shepherd (1974), with SP and DP superficial and deep pyramidal cells; I intrinsic neurons (excitatory stellate cells and inhibitory interneurons lumped together); SA specific sensory afferents; NSA non-specific sensory afferents; rc recurrent collateral. **B** is the scheme of Douglas et al (1989), with GABA inhibitory cells, P2+3 and P5+6 superficial and deep pyramidal cells. **C** is the refinement of **A** in Shepherd (1988), where again SP and DP are superficial and deep pyramidal cells; ST is a stellate cell; SI a superficial inhibitory cell; CH a chandelier cell; BC a basket cell. **D** and **E** are the same circuits of, respectively **A** and **B**, redrawn by Shepherd (2004a, p.37) for an easier comparison.

5.2 Statistical canonical circuits

Douglas and Martin went on by pursuing a different strategy in constructing a more comprehensive microcircuit template of the cortex, using data from more sophisticated experimental methods. Binzegger et al (2004) used three-dimensional cell reconstruction on a sample of primary visual cortex, and analyzed the laminar pattern of synaptic boutons of 39 reconstructed neurons. The average number of synapses formed between neurons in different layers was estimated using an enhanced version of a rule from (Peters and Payne, 1993). In its simplest form this rule states that the synapses from a given type of presynaptic neuron distribute evenly over the population of potential postsynaptic cells in the same cortical layer. In a refined version, additional

details are taken into account, such as that chandelier cells form synapses with pyramidal cells only. The final result is not a circuit any more, but rather a graph of synaptic connections between every type of cell, in five layers (layer II and III are joined together), with the estimated proportion of synapses as the edges.

Their analysis used 14 types of cells, and in the resulting graph each receives synapses on average from 12 other cell types, making the overall graph difficult to read. Even if only one or two cell types form the majority of synapses on the target cell, the remaining connections cannot be assumed negligible, as demonstrated by the thalamic projections, which amount to only 5% but are clearly of fundamental importance. This work is part of a new and rapidly growing field in neuroscience, the study of properties of neural connections from the perspective of graph theory (Bullmore and Sporns, 2009; Hagmann et al, 2010).

A different way of deriving a statistical canonical circuit of the cortex is by using, instead of cell morphology, cellular recordings. Thomson et al (2002) used paired recording, the simultaneous continuous measurement of electrical potentials from presynaptic and postsynaptic sites, obtaining about 1000 recordings on a variety of cortical neurons in several layers. Haeusler and Maass (2007) used these data to assemble a statistical circuit made of 6 virtual cell types, corresponding to excitatory or inhibitory populations of cells distributed into layers II/III, IV, and V. This graph can include two types of edges: probability of connections, as in Binzegger et al., but also average strengths of connections.

Haeusler & Maass took up the burden of showing a computational advantage for the proposed canonical structure, even if established only by a statistical graph. They implemented a network of about 500 single-compartment neurons, with the proportion of connections matching those of the graph. They then performed a series of low-level computational tasks, such as classifying two different sequences of spikes. The performance was compared with the same task executed by networks with the same number and type of neurons, but without the layered structure and the proportions of connections derived from real cortical data.

The results of these analyses do suggest that the observed organization is important. The model derived from the statistical canonical graph outperforms an "amorphous" comparison model, with all connections scrambled in a random way, a result that is perhaps not too surprising. Haeusler et al (2009) also implemented within the same neural simulation, the statistical graph of Binzegger et al., and found very similar performance. However, the performance of the canonical models were almost equal to a control model, where the random scrambling of connections preserved the degrees of the nodes. It is interesting to highlight that the motivations expressed by Haeusler and his collaborators for their statistical canonical models is again in relation to the U/V issue:

> Neurobiological studies have shown that cortical circuits have a distinctive modular and laminar structure, with stereotypical connections between neurons that are repeated throughout many cortical areas. It has been conjectured that these stereotypical canonical microcircuits are [...] advantageous for generic computational operations that are carried out throughout the neocortex
>
> (Haeusler et al, 2009, p.73)

A different simulator has been developed (Potjans and Diesmann, 2014) based on the combined data of Binzegger et al. and Thomson et al., which has better accuracy in predicting certain experimental findings, such as spontaneous firing rates, but performance on computational tasks has not yet been evaluated.

Arguably the very detailed cortical column simulations in the Blue Brain Project (Markram, 2006) and the subsequent (HBP) Human Brain Project (Markram et al, 2015) are the logical outcome of these statistical models. Within subproject SP6 of the HBP, *Brain Simulation* the microcircuitry based on statistical measures reproduces a volume of 0.3 mm^3 of the rat somatosensory cortex with 31 thousands neurons and 37 million synapses, the largest and more detailed ever achieved. This microcircuitry is able to reproduce activities and several response properties recorded in vitro and in vivo experiments. Despite this success, the significance of such simulations in progressing our understanding of the cortex has been seriously questioned Abbott (2014); Eliasmith and Trujillo (2014). In the view expressed in this paper, the main limiting factor of large-scale simulations like HBP is precisely in their derivation from canonical microcircuits: no room for development. Among the 13 subprojects inside HBP only in SP4, *Theoretical Neuroscience*, there is one workpackage (WP4.3) that aims at formulating synaptic plasticity algorithms, but there is no integration whatsoever between this workpackage and the mainstream subproject SP6. Without rules that introduce neurally plausible synaptic development, there is no way to explain U/V: the reconstructed circuit can only reproduce a single, specific, portion of the cortex.

5.3 What canonical microcircuits have explained

Before discussing in detail the consequences of not including development aspects, it is only fair to begin by discussing what canonical microcircuits have instead achieved. In the eyes of the proponents of canonical circuits, one of the major achievements was a fundamental shift away from the "neural doctrine", the neuron-centric view of the brain, legacy of Santiago Ramón y Cajal. Indeed, Shepherd (1991) was so fond of the value and merits of the neural doctrine that he wrote a book on its history *Foundations of the neuron doctrine*. However, in the 25th anniversary edition, the chapter *Modern Revisions of the Neuron Doctrine* warned that the neuron as a whole could be regarded as a unit at one level of organization only, that there are lower levels of functional subunits, and that (Shepherd, 2016, p.289) "above it are several higher levels of multineuronal units. The formation of these different levels is to a large

extent an expression of the complexity of 'local circuits' within each region". The departing from the neural doctrine is described in more radical terms by Douglas and Martin (1991, p.286):

> Now there are signs from experimental and theoretical work on the neocortex that we are on the threshold of a revolution in which the hegemony of the single neuron will be replaced by much more circuit–oriented concepts.

The view of canonical circuits as a scientific revolution (in Kuhn's sense) against the neural doctrine still finds supporters today (Casanova, 2013). In the light of the amount of progress that has been done on canonical circuits after 1990, and their impact on neuroscience as a whole, Shepherd's judgment seems more balanced: canonical circuits did not supersede the neural doctrine, rather they enforced a level of analysis next, and parallel, to that of the neuron as a whole. Multi-level analysis is a requirement that is today largely acknowledged in neuroscience (Craver, 2007) and cognitive neuroscience (Boone and Piccinini, 2016).

There is a list of specific phenomena pertaining to this level of analysis, for which the proposed canonical circuit seems to offer adequate explanations. One is the persistence of excitation and inhibition far longer than the synaptic delays of the units: activation of the cortex sets in motion a sequence of excitation and inhibition in every neuron that self-sustains for some time. Another, is the input amplification by recurrent intracortical connections, so that the thalamic input does not provide the major excitation of any neuron, which is instead due to intracortical excitatory connections. Later on, Douglas and Martin (2004) demonstrated the way in which relatively small changes of inhibitory feedback in the canonical circuit may cause large changes in the gain of the system.

How complete and veridical the explanations of these phenomena are by way of the canonical circuit can be assessed using the 3M criteria of Kaplan. If we apply a strict interpretation of the 3M criteria, we find that none of the models pass it entirely. Shepherd's model is mainly construed as a descriptive circuit, composed of elements that approximately map onto biological parts. It also includes at a descriptive level the nature of the activities of the elements and their relations (i.e. excitatory or inhibitory actions). However, there is no corresponding (additional) computational description, with functional variables that can be related to these components, or equations that match dependencies posited among components of the target mechanism. Only when paired with a computational account, can an answer be given to why the proposed microcircuit subserves the enormous computational and cognitive capabilities of the cortex. Douglas and Martin's models do include simplified computational accounts of the relations between the elements in the circuit, namely the dependency of the firing rate of the units upon the magnitude of their excitation or inhibition. These dependencies are supposed to correspond to the causal relations among the physical cells in the cortex.

However, in all models the elements are idealized in terms of *population*: element P1 in Shepherd's model stands for a population of superficial pyramidal

cells that, within a small enough cortical area, cooperate to process the same information. In the case of Douglas et al., P1 represents populations of superficial pyramidal cells together with the pool of spiny stellate cells that project to them. Therefore, constraint (a) of 3M, "the variables in the model correspond to *identifiable* components [...] of the target mechanism", is not met. There is an ontological gap that prevents a mapping between populations of units and individual units. Note that this is very different from the issue of the amount of details included in a model with respect to the target mechanism. For example one may well abstract the output of a single neuron with firing rate coding, or the transmission of a synapse with a scalar quantity proportional to its efficiency. Much detail on the working of the neuron and the synapse will be missed, but it would still be possible to establish correspondence between that neuron and that synapse and identifiable components of a target neural mechanism. In the case of canonical circuits, units are clearly not physical single cells, their extension in the cortex is not specified, nor the number and locations of cells, on which the population of cells is averaged as a single abstract unit. As highlighted by Craver (2007, p.131) mechanistic explanation is inherently componential, and distinguishing *how-possibly* from *how-actually* models, requires that one distinguish real components from fictional posits. Components in canonical models are fictions, even if useful fictions that quite likely capture much about populations of cells.

From this point of view, the efforts towards statistical canonical circuits help in specifying what mapping can be established between the mesoscopic components in the canonical circuit and parts in the cortex. The graph underlying statistical canonical circuits by itself, is not a computational model anymore and therefore, cannot be evaluated against Kaplan's 3M constraint, however, it is an approach defended by Bechtel (2015) as a generalized mechanistic explanation. When the number of components of a system is large but repetitive, Bechtel et al. argue that graph theoretical analysis of its organization can contribute to the explanation of why mechanisms implementing this form of organization exhibit the behavior they do. One of the distinctive features of graph analysis is the classification of networks inside a well established taxonomy (Watts and Strogatz, 1998). This is, for example, what Binzegger et al (2009) provide for their statistical canonical circuit, resulting in the observation that the topology of the graph changes from a random graph structure, when the fraction of simultaneously active neurons is low, to a small-world structure when high. However, there is no obvious and direct relation between the characterization of cortical circuits in terms of graph topology, and their computational power.

5.4 What canonical microcircuits have missed

The explanation primarily missed by these circuits is U/V, which Martin considered to be the $64000 question (see sec. 5.1). Each of the canonical circuits reviewed here seem to perform a single, fixed operation given its inputs, and

demonstrating that the same circuit could do useful information processing across disparate modalities would be a significant (and surprising) additional achievement. The existing proposed canonical circuits focus on explaining a subset of *shared* properties of cortical regions.

The brief and tentative suggestions on how canonical circuits may cope with diversity found in Shepherd, and in Douglas & Martin, are quite different. For Shepherd (2004a, p.35) "An objection to the idea of a regional basic circuit is that it does not adequately represent the rich diversity of neural elements and synaptic connections that can be found in most brain regions. The basic regional circuit can be expanded with subcircuits for specific functions as needed". Instead, for Martin (1988, p.640) "the only thing that would distinguish auditory cortex from motor cortex is that it has different sources of input and sends its output to different targets", and Douglas and Martin (2004, p.420) "it might reasonably be argued that the similarities and regularities are incidental to the many different mappings of input to output that are evident in the different neocortical areas".
The position of Shepherd can be schematized as follows:

$\mathcal{P}_1$ the uniformity of the cortex is reflected in a canonical circuit;

$\mathcal{P}_2^S$ the variety of functions in the cortex can be accounted by the integration of the canonical circuit with subcircuits.

And here that of Douglas & Martin:

$\mathcal{P}_1$ the uniformity of the cortex is reflected in a canonical circuit;

$\mathcal{P}_2^{DM}$ the variety of functions in the cortex can be accounted by the different inputs and outputs.

Of course propositions $\mathcal{P}_2^S$ and $\mathcal{P}_2^{DM}$ are just hypotheses, that in turn would require descriptions of the mechanisms that lump together the canonical circuit with the "different mappings of input to output" (for $\mathcal{P}_2^{DM}$) or "subcircuits" (for $\mathcal{P}_2^S$). But even before that, the two hypotheses fall short of explaining empirical data, one for all the famous rewiring experiments, in which retinal axons of ferrets are connected at birth to the medial geniculate nucleus, which relays the signals to A1 instead of V1. This abnormal connectivity induced a functional reorganization of A1, that enabled visual behavior in the animals (Roe et al, 1987; Sur, 1989; Roe et al, 1990). This result argues in favor of proposition $\mathcal{P}_1$: if A1 adapted in order to work similarly to V1, it is plausible that the two cortical areas share a common basic circuit, and Sur (1989) cite specifically Shepherd (p.46): "The notion that different parts of sensory thalamus or neocortex share basic commonalities is not new (Lorente de No 1938: Mountcastle 1978: Shepherd 1979)." But both propositions $\mathcal{P}_2^S$ and $\mathcal{P}_2^{DM}$ are invalidated: there is a deep change in the organization across a major tonotopic axis of A1, into a periodical, symmetrical array of orientation-tuned clusters of neurons, resembling that of V1 (Gao and Pallas, 1999), and these changes are assessed at adulthood only.

I should add that Douglas & Martin were perfectly aware that hypothesis $\mathcal{P}_2^{DM}$ was just tentative, and explaining diversity might become awkward (Douglas and Martin, 1991, p.291-292):

> [...] it seems perverse to regard the visual cortex as an *ad hoc* collection of specialist circuits, rather than a set of basic circuits adapted to perform many different tasks. Nevertheless, we should certainly not expect the present course of experimental and theoretical work to provide us with any early answers as to how neurons become orientation selective, or direction selective. Although the solution of any one of these problems would be very impressive, given the trail of failed attempts, it is also not self evident that we would thereby arrive at any general rules about the operation of the cortical circuits.

The rewiring experiment brings us to the main reason, in my opinion, for the scarce explanatory power of canonical circuits in resolving the U/V issue. It is the dismissal of Marr's fundamental issue of how the cortex can *learn* to perform a wide variety of tasks. Sur (1989, p.46) gave a suggestion in this direction, with respect to the results of the rewiring experiment:

> We suggest that, in auditory cortex of operated animals, the response features that depend specifically on the two-dimensional nature of visual input indicate a form of adaptive self-organization in cortex. [...] The mechanism behind such organization or adaptation might generally involve spatiotemporal coactivation in subsets of the visual input along with lateral inhibition, enabling modification of synaptic efficacy between presynaptic elements and restricted groups of postsynaptic neurons or sets of postsynaptic elements.

There are historical justifications for the initial discarding of development: in the period when the first canonical microcircuits were proposed the picture on cortical plasticity was certainly scattered and much more obscure than today. One of the fundamental papers organizing the state of the art in cortical plasticity appeared in 1998 Buonomano and Merzenich (1998), and it is in the same period that STDP (Spike-Time-Dependent-Plasticity), the main form of cortical plasticity, was discovered Song and Abbott (2001). Nevertheless, more that ten years earlier von der Malsburg (1973) had proposed the first mathematical simulation of the development of orientation sensitive cells in the striate cortex by self-organization. Since then, the simulation of cortical development by self-organization become a well established research direction Miikkulainen et al (2005); Willshaw (2006), but no fruitful interaction with the canonical circuit research ever took place.

All canonical microcircuits address only a single static adult configuration, discarding the emergence of specific synaptic connections driven by the specific input patterns actually encountered by a cortical region during development. Even when models with correspondence between units and statistics of real cortex can be simulated (as in Haeusler & Maass), all synaptic strengths are necessarily equal to the statistical average derived by the data. Yet if the synaptic strengths in the circuit corresponding to one orientation-selective column in the primary cortex were all substituted with their mean value, the column

would lose its orientation selectivity, thus entirely missing its computational function.

Already in 2002, during a discussion on universal cortical circuits and functional diversity that gathered 48 neuroscientists in Madrid, reported by Nelson (2002b), the issue of plasticity was raised (p.23):

> Ultimately, studying cortical microcircuit structure is valuable only in what it can tell us about cortical function. [...] How do perceptual, motor, and cognitive abilities arise out of complex patterns of neuronal activity across the cortex? [...] The problem is compounded by the fact that cellular and synaptic dynamics not only influence activity; they are themselves plastic in an activity–dependent manner.

The need to integrate development into a canonical core of the cortex has been recently raised by Douglas and Martin themselves, in their comments to a target article by DeFelipe (2015) about the difficulties in studying the brain by connectomics (DeFelipe et al, 2016, p.2-3):

> Our line of reasoning suggests that a much more profitable approach is not brute force dense reconstruction of partial circuits in individual animals, but instead to identify the principles of connectivity from the point of view of a self–constructing connectome [...]. The development of neocortex is particularly interesting in this regard, because the regularities of structure discussed above raise questions of how apparently complex neuronal connectivity could be established on a basis of relatively simple developmental rules enclosed in only a few precursor cells.

The new effort in the identification of the "relatively simple developmental rules" for cortical circuits has begun, and the Institute für Neuroinformatick of Zurich is doing its part (Bauer et al, 2014), but is still in its early infancy. According to Wright and Bourke (2017) one of the major hindrances for a unified theory of cortical development is the uncertainty regarding columnar cortical organization, or the lack of it, which we discussed in section 3.3.

6 Canonical computations

While the canonical models presented in the previous section included few (if any) computational explanations, others have taken an alternative direction. These models instead give precedence to mathematical formulations, and only then derive, as a subordinate effort, neural circuits potentially able to implement the formulations.

This approach is quite different from that of Marr, who developed a mathematical framework as a pure speculation of the general problem of event classification. In the models discussed in this section, mathematical operations have been identified on sound empirical grounds, by careful analysis of which computations are most often carried out across different areas in the cortex. The important shift with respect to the perspective of Sherrington and Douglas is in assigning the "canonical" character to *computations* rather than circuits: it is the general applicability of certain operations that makes

Table 1 Different cortical computations obtained from equation (1) with several exponents.

operation	p	q	r
energy model	2	2	0
divisive normalization	2	2	1
Gaussian-like	1	2	1
max-like	3	2	1

the cortex powerful and flexible, operations that could even be carried out by different circuits in different cortical areas.

The most influential work in this stream of research is that of Kouh and Poggio (2008). They proposed the following response y to a set of n inputs x_i as "canonical":

$$y = \frac{\sum_{i=1}^{n} w_i x_i^p}{k + \left(\sum_{i=1}^{n} x_i^q\right)^r} \tag{1}$$

The inputs x_i can be regarded as activation of neurons that project to a single neuron, whose activation is the value of variable y. The efficiency of the synaptic connections are the values of weights w_i. By changing the exponents, this equation can diversify to assume a number of shapes typically observed as cortical responses, summarized in Table 1. Note that the exponents p, p, r, and the parameter k do not bear any correspondence with biological entities or quantities in the neural system.

Kouh and Poggio argue that this set of nonlinear operations provides a broad coverage of proposed cortical computations. The energy model has been used in explaining the phase invariance of complex cells in the primary visual cortex, derived by half-squaring and summing the responses of the quadrature pairs (Adelson and Bergen, 1985). Divisive normalization has been assumed in various contexts, such as direction selectivity of simple cells in the primary visual cortex to drifting gratings (Heeger, 1993), or competitive attention in visual areas V2 and V4 (Reynolds et al, 1999). Gaussian-like operations are even more widespread, being observed in various features of the primary visual cortex, like orientation, spatial frequency, direction, and velocity tuning (Rose, 1979; Daugman, 1980; Grzywacz and Yuille, 1990), and in the frequency responses of some cells in the nonprimary auditory cortex (Rauschecker et al, 1995). There is less direct evidence of a max-like operation in cortical circuits, but there is support from physiological experiments (see citations in Kouh and Poggio, 2008).

In addition to this mathematical framework, Kouh and Poggio proposed a circuit sketch, in which a single abstract neuron computes the response y from three inputs x_i, and a shunting inhibition of their pooled values, along with a demonstration that a compartment model can be compatible with this sketch, using certain values of its parameters. This model, of course, has no direct mapping with cell types and layers in the cortex, as in the canonical circuits seen in §5.1, it is just the scheme for a single neuron. Despite the fundamental

difference in emphasis, Kouh and Poggio still used the expression "canonical neural circuit" in the title of their paper.

A few years later Carandini and Heeger (2012) abandoned the idea of a circuit altogether, focusing instead on the search for "canonical neural computation". Carandini and Heeger proposed that divisive normalization is such a computation, reviewing a comprehensive set of evidence for such operations in the primary visual cortex and in other areas such as the visual motor area MT and the lateral intraparietal area LIP. Despite the cardinal role suggested for divisive normalization, they did not argue that it is *the* unique canonical computation, but rather one in a small set, together with at least exponentiation and linear filtering. Their formulation of divisive normalization is slightly different from that in equation (1), but using the variable-name convention from Kouh and Poggio, it has the following expression:

$$y = \gamma \frac{x_0^p}{k^p + \sum_{i=1}^{n} x_i^p} \tag{2}$$

with just a single exponent p, and with x_0 the driving input, which is assumed to be a weighted sum of multiple signals.

The paper lists many different applications of normalization: maximizing sensitivity to changes in input, achieving invariance with respect to some stimulus dimension, perform max-pooling operation, reduce redundancy by incrementing the statistical independence of responses. At the same time, normalization is not related to a specific neural mechanism (p.58):

> [...] it is unlikely that a single mechanistic explanation will hold across all systems and species: what seems to be common is not necessarily the biophysical mechanism but rather the computation. Moreover, in some systems ([...]) normalization seems to result from multiple circuits and mechanisms operating in concert and cascading across multiple stages.

The pervasiveness of normalization across different informational demands and across different neural circuits called for several philosophical reflections. For Marcus et al (2014) canonical computations may sound the death knell for canonical circuits (p.551–552):

> there is still no consensus about whether such a canonical circuit exists, [...] Likewise, there is little evidence that such uniform architectures can capture the diversity of cortical function [...] One possibility is that neuroscience's query should not be a single canonical circuit, but a broad array of reusable computational primitives [...] Candidate computational primitives might include circuits [...] for normalizing the ratio between the activity of an individual neuron and a set of neurons.

6.1 Where canonical computations come from

The major drawback of canonical computations, in my view, is similar to that observed for the proposed canonical circuits: that of overlooking development,

which was the original focus of Marr's earliest attempt. Equations (1) and (2) are static, describing fixed aspects of responses of neurons or cortical neural aggregations. They say nothing of the mechanisms that lead neurons and cortical circuits to perform functions, in which the properties described by those equations can be identified. For instance, why do primary visual cortex neurons become selective for Gabor patterns while neurons in other modalities might become selective for quite different shapes? There is a number of models that attempt to simulate, for example, how Gabor responses may arise in cortical structures by natural development (Miikkulainen et al, 2005; Lücke, 2009), but outside of canonical computation concepts. Although it is theoretically conceivable that a static fixed set of computations could account for the diverse behavior of neurons across the cortex, such a strong claim would require extraordinary evidence, and current computational proposals come nowhere near accounting for this diversity. In fact Kouh and Poggio (2008) touch the issue, and provide, in Appendix C of their paper, a brief outline of how Gaussian-like and max-like operations can be learned. However, the aim of their Appendix C is simply to prove the concept that Gaussian-like and max-like operations can, in principle, be learned. Therefore, differently from the careful comparison with empirical data done when identifying the main canonical computation, for the purpose of demonstrating their possible development an abstract supervised scheme, void of biological plausibility, is instead used. It was clearly beyond the scope of their work, to investigate in detail how different canonical computations may develop.

6.2 Cicadas and exponents

Divisive normalization as canonical computation has been taken by Chirimuuta (2014) as a case study, along with Gabor functions, in endorsing the distinctness of computational explanations from mechanistic explanations, and here this claim is discussed. To explore the idea of whether a computational explanation could be sufficient, let us consider a phenomenon in biology where the explanatory power of a mathematical explanation appears compellingly high: the famous case of the life-cycle of some North American cicadas, analyzed by Baker (2005). In three species of genus *Magicicada* the nymphal stage remains in the soil for a period of either 13 or 17 years, depending on the geographical area. Then the adult cicadas emerge synchronously, and within the same few days mate and then die a few weeks later. The cycle then repeats itself. For these insects, it is crucial to have sufficient mating opportunities during their brief adult stage, which is the reason for synchronizing, but it is almost as important to avoid mating with subspecies that have different cycle periods. Mating between subspecies would produce offspring that would not be coordinated with either subspecies. There are also biological constraints that limit the periods to be within 20 years.

The intersections of two periodic numerical series are minimized if the two periods are *coprime* (they have no common factors other than 1). Thus the ap-

parently weird periods of 13 and 17 for neighboring cicadas can be explained on the basis of number-theoretic results. Baker argues that the number-theoretic theorem is genuinely explanatory in its own right, in that it explains why prime periods are evolutionarily advantageous for cicadas. Of course, not everyone is convinced; philosophers that argue that mathematics cannot explain any natural phenomena, specifically challenge this example as well (Daly and Langford, 2009). Here for the sake of argument we simply assume that there is no case more convincing that the cicada example for the indispensability of a mathematical argument as the fundamental explanation (Lange, 2013), and consider whether canonical cortical computations could reach this standard.

So, are there any mathematical facts in equations (1) or (2), for example the exponents that play a crucial role in shaping the functions, that are genuinely explanatory regarding phenomena of the cortex, with substantial evidence that is similar in weight to that of the prime numbers for cicadas? Clearly, not in any obvious way. It is elegant and convenient to unify in the single equation (1) all the cases listed in Tab. 1, by assigning a few integer values to the exponents, but unlike the coprime numbers in the case of Baker, there are no mathematical properties of the exponent values that appear indispensable for the explanation of the phenomenon. As stated by Carandini and Heeger (2012, p.54) "The constants γ, k and p [in our notation] constitute free parameters that are typically fit to empirical measurements", and, for example, they report a range for p in neurons of the primary visual cortex from 1.0 to 3.5.

Chirimuuta (2014) suggests that the Gabor model for the receptive fields of simple cells in the primary visual cortex is additional evidence for the autonomy of mathematical explanations in neuroscience. Functions in this family can be shown to minimize the joint uncertainty over time and frequency of signals (Gabor, 1946; Daugman, 1985), so it can be argued that receptive fields take the form of Gabor functions because they follow the *efficient coding principle*, which she argues also applies to divisive normalization.

One general objection might be that brain computations rarely obey abstract principles of absolute optimization, for natural neurophysiological constraints. Above all, while the Gabor case shares similarity with cicadas in having a precise theoretical property that meets with biological requirements, normalization does not follow directly from an efficient coding principle. In fact, in a recent paper Chirimuuta (2017) details her defense of explanations in neuroscience that are distinctively mathematical, according to the causal/non-causal criteria of Lange (2013), only for the Gabor case, without attempting to include canonical computations. Paz (2017) argue that canonical neural computations are not valid counter-examples for the mechanistic framework of explanation in computational neuroscience, since only according to interpretations that include the causal organization underlying a phenomenon, are canonical computations explanatory.

Our point here is that canonical computations not only cannot stand as autonomous mathematical explanations like in the cicada case, but as Paz has suggested, are not explanations at all, when taken as abstract computations. Normalization is widespread, but unlike Gabor receptive fields, it is often a side

component inside a variety of response functions. In several specific phenomena, like cross-orientation suppression in V1 (Priebe and Ferster, 2006) there is a role for divisive normalization within the overall computation, but the explanation of the phenomena becomes adequate when parts, activities, and organizational features of the mechanism in V1 underlying cross-orientation suppression are described, in other words, when 3M constraints are met.

The point made is not detrimental to the scientific value of canonical computations as general tools, with no explanatory power by themselves. The identification of components inside an overall computational function, and the assignment of components to established mathematical classes such as divisive normalization, is of immense theoretical value and helpful in making progress in the understanding of cortical computations. However, we doubt that canonical computations can be considered or cited as being the novel replacements of canonical circuits, as assumed by (Marcus et al, 2014). Their position closely resembles unificationist accounts of explanation (Kitcher, 1981), implying that the highest explanatory power is achieved when common abstract computations uniformly capture a variety of tasks performed across multiple neural circuits. But the idea that more unifying models should be explanatorily superior to those that are less unifying is controversial at the very least (Sober, 1999), and is at odds with the mechanistic framework of explanation in neuroscience (Kaplan, in press).

7 Conclusions

This paper analyzed the concept of a "canonical" core for the cortex, that opened up a new domain of inquiry in the 1970's. This research seems largely motivated by questions regarding how the cortex can simultaneously be so uniform in anatomical structure and yet so diversified in its functions. This search continues to be meaningful, because there is evidence for at least a rough uniformity of cortical structure, and strong evidence for functional diversity. The earliest approach, that of David Marr, followed a framework resembling deductive-nomological explanations, starting from a mathematical account of the general problem of event classification, deriving relevant theorems, and finally searching putative correspondences between variables in the mathematical framework and cortical components. This approach had little hopes to uncover a canonical core for the cortex, but had among its merits its focus on the developmental aspects of cortical circuits. Approaches in the 80s and 90s, that can be framed as mechanistic, have been successful in outlining cortical microcircuits that may play some canonical role, and are able to reproduce certain features of cortical computations. However, they have all failed to derive a computational account of the circuits in any way powerful enough to fully explain the performance of the cortex, a failure that I argue was implicit in the lack of a neural development component. At the beginning of this century the search for a canonical microcircuit took new directions, trying to incorporate data from more sophisticated experimental methods. These statistical

canonical circuits can be framed into a generalized mechanistic explanation, where the identification of components is blurred, and integrated by graph theoretical analysis. During the last two decades alternative canonical solutions have been proposed, giving precedence to mathematical formulations, and only then deriving neural circuits potentially able to implement the formulations. Putative canonical computations have been theoretically defended as valid computational explanations, independently from the identification of possible mechanisms. Canonical computations, like divisive normalization, offer important insights, yet do not bring us any closer to answering the main question of the cortex, first of all due to the difficulty of relating computations to actual cortical architecture. The computations collected under the umbrella of canonical computations do not seem to have explanatory power, when taken in isolation, but are rather components that may play a role when integrated in an overall cortical function. Even more importantly, canonical computations suffer the same limitation of canonical circuits in overlooking the crucial dimension of cortical development, that was assumed by David Marr as one of the primary issues in searching for a canonical core of the cortex. This drawback seems acknowledged by a number of researchers today. The integration of circuital and developmental concepts may uncover promising research toward an understanding of the cortex, and in particular, of the perplexing clash between the uniformity and the variety of its functions.

References

Abbott A (2014) Row hits flagship brain plan. Nature 511:133–134

Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America 2:284–299

Anzai A, Peng X, Van Essen DC (2007) Neurons in monkey visual area V2 encode combinations of orientations. Nature Neuroscience 10:1313–1321

Baker A (2005) Are there genuine mathematical explanations of physical phenomena? Mind 114:223–238

Bauer R, Zubler F, Pfister S, Hauri A, Pfeiffer M, Muir DR, Douglas RJ (2014) Developmental self-construction and -configuration of functional neocortical neuronal networks. PLoS Computational Biology 10:e1003,994

Bechtel W (1982) Two common errors in explaining biological and psychological phenomena. Philosophy of Science 49:549–574

Bechtel W (2015) Generalizing mechanistic explanations using graph-theoretic representations. In: Braillard and Malaterre (2015a), pp 199–225

Bergeron V (2007) Anatomical and functional modularity in cognitive science: Shifting the focus. Philosophical Psychology 20:175–195

Berlin R (1858) Beitrag zur structurlehre der grosshirnwindungen. PhD thesis, Medicinischen Fakultät zu Erlangen

Betz VA (1874) Anatomischer Nachweis zweier Gehirncentra. Zentralblatt medicinischen Wissenschaften 12:578–599

Binzegger T, Douglas RJ, Martin KA (2004) A quantitative map of the circuit of cat primary visual cortex. Journal of Neuroscience 24:8441–8453

Binzegger T, Douglas RJ, Martin KA (2009) Topology and dynamics of the canonical circuit of cat V1. Neural Networks 22:1071–1078

Boone W, Piccinini G (2016) The cognitive neuroscience revolution. Synthese 193:1509–1534

Borck C (2001) Electricity as a medium of psychic life: Electrotechnological adventures into psychodiagnosis in weimar germany. Science in Context 14:565–590

Bower J (2005) Looking for newton: Realistic modeling in modern biology. Brains, Minds, Media 1:bmm217

Braak H (1974) On the structure of the human archicortex. i. the cornu ammonis. a Golgi and pigment architectonic study. Cell Tissue Research 152:349–383

Braillard PA, Malaterre C (eds) (2015a) Explanation in Biology – An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences. Springer-Verlag, Berlin

Braillard PA, Malaterre C (2015b) Explanation in biology: An introduction. In: Braillard and Malaterre (2015a), pp 1–28

Brazier M (1961) A history of the electrical activity of the brain: The first half-century. Macmillan, New York

Brindley GS, Lewin W (1968) The sensations produced by electrical stimulation of the visual cortex. Journal of Neurophysiology 196:479–493

Brodmann K (1903) Beiträge zur histologischen lokalisation der Grosshirnrinde, 1: die Regio Rolandica; 2: der Calcarintypus. Journal Psychol Neurol 2:79–107,133–159

Brodmann K (1909) Vergleichende Lokalisationslehre der Grosshirmrinde. Barth, Leipzig

Brown AG, Fyffe R, Noble R, Rose P, Snow P (1980) The density, distribution and topographical organization of spinocervical tract neurones in the cat. Journal of Physiology 300:409–428

Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience 10:186–198

Buonomano DV, Merzenich MM (1998) Cortical plasticity: From synapses to maps. Annual Review of Neuroscience 21:149–186

Burnston DC (2016a) Computational neuroscience and localized neural function. Synthese 193:3741–3762

Burnston DC (2016b) A contextualist approach to functional localization in the brain. Biology and Philosophy 31:527–550

Buxhoeveden DP, Casanova MF (2002) The minicolumn hypothesis in neuroscience. Brain 125:935–951

Carandini M, Heeger D (2012) Normalization as a canonical neural computation. Nature Reviews Neuroscience 13:51–62

Carlo CN, Stevens CF (2013) Structural uniformity of neocortex, revisited. Proceedings of the Natural Academy of Science USA 110:719–725

Casanova MF (2013) Canonical circuits of the cerebral cortex as enablers of neuroprosthetics. Journal of Experimental Social Psychology 49:719–725
Casanova MF, Opris I (eds) (2015) Recent Advances on the Modular Organization of the Cortex. Springer-Verlag, Berlin
Chalupa L, Werner J (eds) (2003) The Visual Neurosciences. MIT Press, Cambridge (MA)
Chirimuuta M (2014) Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. Synthese 191:127–153
Chirimuuta M (2017) Explanation in computational neuroscience: Causal and non-causal. British Journal for the Philosophy of Science axw034
Church A (1941) The Calculi of Lambda Conversion. Princeton University Press, Princeton (NJ)
Churchland PS, Sejnowski T (1994) The Computational Brain. MIT Press, Cambridge (MA)
Copeland JB, Posy CJ, Shagrir O (eds) (2013) Computability: Turing, Gödel, Church, and Beyond. MIT Press, Cambridge (MA)
Cowan JD (1991) Commentary on *a theory for cerebral neocortex*. In: Vaina L (ed) From the Retina to the Neocortex, Birkhäuser, Boston, pp 203–209
Craver CF (2001) Role functions, mechanisms, and hierarchy. Philosophy of Science 68:53–74
Craver CF (2007) Explaining the brain: mechanisms and the mosaic unity of neuroscience. Oxford University Press, Oxford (UK)
Craver CF, Darden L (2013) In Search of Mechanisms: Discoveries across the Life Sciences. University of Chicago Press, Chicago (IL)
Cummins R (1975) Functional analysis. Journal of Philosophy 72:741–765
Daly C, Langford S (2009) Mathematical explanation and indispensability arguments. Philosophical Quarterly 59:641–658
Daugman JG (1980) Two-dimensional spectral analysis of cortical receptive field profiles. Vision Research 20:847–856
Daugman JG (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Society of America 2:1160–1169
Dayan P, Abbott LF (2001) Theoretical Neuroscience. MIT Press, Cambridge (MA)
DeFelipe J (2015) The anatomical problem posed by brain complexity and size: a potential solution. Frontiers in Neuroanatomy 9:Article 104
DeFelipe J, Douglas RJ, Hill SL, Lein ES, Martin KAC, Rockland KS, Segev I, Shepherd GM, Tamás G (2016) Comments and general discussion on "the anatomical problem posed by brain complexity and size: A potential solution". Frontiers in Neuroanatomy 10:Article 60
Dehaene S (2014) Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Viking Adult, New York
Douglas RJ, Martin KA (1991) Opening the grey box. Trends in Neuroscience 14:286–293

Douglas RJ, Martin KA (1992) In search of the canonical microcircuits of neocortex. In: Lent R (ed) The Visual System from Genesis to Maturity, Springer-Verlag, Berlin, pp 213–232

Douglas RJ, Martin KA (2004) Neuronal circuits of the neocortex. Annual Review of Neuroscience 27:419–451

Douglas RJ, Martin KA, Whitteridge D (1989) A canonical microcircuit for neocortex. Neural Computation 1:480–488

Douglas RJ, Mahowald MA, Martin KAC (1994) Hybrid analog-digital architectures for neuromorphic systems. In: Proceedings of the IEEE International Conference on Neural Networks, IEEE, pp 1848–1853

Douglas RJ, Markram H, Martin K (2004) Neocortex. In: Shepherd (2004b), pp 499–558, 5th Edition

du Bois-Reymond E (1849) Untersuchungen über Thierische Elektricität. G. Reimer, Berlin

Eliasmith C, Trujillo O (2014) The use and abuse of large-scale brain models. Current Opinion in Neurobiology 25:1–6

Farhat NH (2007) Corticonic models of brain mechanisms underlying cognition and intelligence. Physics of Life Reviews 4:223–252

Fodor J (1975) The Language of Thought. Harvard University Press, Cambridge (MA)

Fresco N (2014) Physical Computation and Cognitive Science. Springer-Verlag, Berlin

Fuster JM (2008) The Prefrontal Cortex. Academic Press, New York, fourth edition

Gabor D (1946) Theory of communication. Journal IEE 93:429–459

Gao WJ, Pallas S (1999) Cross-modal reorganization of horizontal connectivity in auditory cortex without altering thalamocortical projections. Journal of Neuroscience 19:7940–7950

Garson J (2016) A Critical Overview of Biological Functions. Springer-Verlag, Berlin

Glennan S (1996) Mechanisms and the nature of causation. Erkenntnis 44:49–71

Grzywacz NM, Yuille AL (1990) A model for the estimate of local image velocity by cells in the visual cortex. Proceedings of the Royal Society of London B 239:129–161

Haberly LB, Shepherd GM (1973) Current-density analysis of summed evoked potentials in opossum prepyriform cortex. Journal of Neurophysiology 36:789–802

Haeusler S, Maass W (2007) A statistical analysis of information-processing properties of lamina-specific cortical microcircuit models. Cerebral Cortex 17:149–162

Haeusler S, Schuch K, Maass W (2009) Motif distribution, dynamical properties, and computational performance of two data-based cortical microcircuit templates. Journal of Physiology – Paris 21:1229–1243

Hagmann P, Cammoun L, Gigandet X, Gerhard S, Grant PE, Wedeen V, Meuli R, Thiran JP, Honey CJ, Sporns O (2010) MR connectomics: Principles and

challenges. Journal of Neuroscience Methods 194:34–45
Harris KD, Shepherd GM (2015) The neocortical circuit: themes and variations. Nature Neuroscience 18:170–181
Haueis P (2016) The life of the cortical column: opening the domain of functional architecture of the cortex (19551981). History and Philosophy of the Life Science 38:2
Heeger DJ (1993) Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. Journal of Neurophysiology 70:1885–1898
Hempel CG (1965) Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. Free Press, New York
Herculano-Houzel S, Collins CE, Wong P, Kaas JH, Lent R (2008) The basic nonuniformity of the cerebral cortex. Proceedings of the Natural Academy of Science USA 34:12,593–12,598
Hines M, Carnevale N (1997) The NEURON simulation environment. Neural Computation 9:1179–1209
Holland O, Husbands P (2011) The origins of British cybernetics: the Ratio Club. Kybernetes 40:110–123
Hoorweg JL (1898) Ueber die elektrischen Eigenschaften der Nerven. Pflügers Arch ges Physiol 71:128–157
Horton JC, Adams DL (2005) The cortical column: a structure without a function. Philosophical transactions of the Royal Society B 360:837–862
Hubel D, Wiesel T (1959) Receptive fields of single neurones in the cat's striate cortex. Journal of Physiology 148:574–591
Hubel D, Wiesel T (1963) Shape and arrangement of columns in cat's striate cortex. Journal of Physiology 165:559–568
Hubel D, Wiesel T (1974) Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. Journal of Comparative Neurology 158:295–305
Jones JP, Palmer LA (1987) The two-dimensional spatial structure of simple receptive fields in cat striate cortex. Journal of Neurophysiology 58:1187–1211
Kaas JH (2012) Evolution of columns, modules, and domains in the neocortex of primates. Proceedings of the Natural Academy of Science USA 109:10,655–10,660
Kaplan DM (2011) Explanation and description in computational neuroscience. Synthese 183:339–373
Kaplan DM (in press) Neural computation, multiple realizability, and the prospects for mechanistic explanation. In: Kaplan DM (ed) Explanation and Integration in Mind and Brain Science, Oxford University Press, Oxford (UK)
Kaplan DM, Craver CF (2011) Towards a mechanistic philosophy of neuroscience. In: French S, Saatsi J (eds) Continuum Companion to the Philosophy of Science, Continuum Press, London, pp 268–292
Karbowski J (2014) Constancy and trade-offs in the neuroanatomical and metabolic design of the cerebral cortex. Frontiers in Neural Circuits 8:9

Kirchhoff G (1845) Ueber den Durchgang eines elektrischen Stromes durch eine Ebene, insbesonere durch eine kreisförmige. Poggendorff's Annalen der Physik und Chemie 64:487–514

Kisvárday Z, Martin KA, Whitteridge D (1985) Synaptic connections of intracellularly filled clutch cells: a type of small basket cell in the visual cortex of the cat. Journal of Comparative Neurology 241:111–137

Kitcher P (1981) Explanatory unification. Philosophy of Science 48:507–531

Klein C (2017) Brain regions as difference-makers. Philosophical Psychology 30:1–20

Kouh M, Poggio T (2008) A canonical neural circuit for cortical nonlinear operations. Neural Computation 20:1427–1451

Lange M (2013) What makes a scientific explanation distinctively mathematical? British Journal for the Philosophy of Science 64:485–511

Lorente de Nó R (1938) Architectonics and structure of the cerebral cortex. In: Fulton J (ed) Physiology of the nervous system, Oxford University Press, Oxford (UK), pp 291–330

Lücke J (2009) Receptive field self-organization in a model of the fine structure in V1 cortical columns. Neural Computation 21:2805–2845

Maçarico da Costa N, Martin KAC (2010) Whose cortical column would that be? Frontiers in Neuroanatomy 4:16

Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. Philosophy of Science 67:1–84

Mahowald M (1994) An Analog VLSI System for Stereoscopic Vision. Kluwer, Dordrecht (NL)

Marcus GF, Marblestone A, Dean T (2014) The atoms of neural computation. Science 346:551–552

Markram H (2006) The blue brain project. Nature Reviews Neuroscience 7:153–160

Markram H, Muller E, Ramaswamy S, et al MWR (2015) Reconstruction and simulation of neocortical microcircuitry. Cell 163:456–492

Marr D (1970) A theory for cerebral neocortex. Proceedings of the Royal Society of London B 176:161–234

Marr D (1975) Review: Approaches to biological information processing. Science 190:875–876

Martin KA, Whitteridge D (1982) The morphology, function and intracortical projections of neurones in area 17 of the cat which receive monosynaptic input from the lateral geniculate nucleus (LGN). Journal of Physiology 328:37–38p

Martin KA, Whitteridge D (1984a) Form, function, and intracortical projections of spiny neurones in the striate visual cortex of the cat. Journal of Physiology 353:463–504

Martin KA, Whitteridge D (1984b) The relationship of receptive field properties to the dendritic shape of neurones in the cat striate cortex. Journal of Physiology 356:291–302

Martin KAC (1988) The Wellcome Prize lecture – from single cells to simple circuits in the cerebral cortex. Quarterly Journal of Experimental Physiology

73:637–702
Martinotti C (1890) Beitrag zum Studium der Hirnrinde und dem Centralursprung der Nerven. Internationale Monatsschrift für Anatomie und Physiologie 7:69–90
Mead C (1989) Analog VLSI and Neural Systems. Addison Wesley, Reading (MA)
Mehta MR, Quirk MC, Wilson MA (2000) Experience-dependent asymmetric shape of hippocampal receptive fields. Neuron 25:707–715
Meynert TH (1869) Studien über die Bedeutung des zweifachen Rückenmarksursprunges aus dem Grosshirn. Sitzungsber Akad Wissensch, II Abt 60:447–462
Miikkulainen R, Bednar J, Choe Y, Sirosh J (2005) Computational maps in the visual cortex. Springer-Science, New York
Miłkowski M (2013) Explaining the Computational Mind. MIT Press, Cambridge (MA)
Miller EK, Freedman DJ, Wallis JD (2002a) The prefrontal cortex: Categories, concepts and cognition. Philosophical Transactions: Biological Sciences 357:1123–1136
Miller LM, Escab MA, Read HL, Schreiner CE (2002b) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. Journal of Neurophysiology 87:516–527
Millikan RG (1989) Biosemantic. Journal of Philosophy 86:242–265
Molnár Z (2013) Cortical columns. In: Rubenstein JLR, Rakic P (eds) Comprehensive developmental neuroscience: neural circuit development and function in the healthy and diseased brain, Academic Press, New York, pp 109–129
Mountcastle V (1957) Modality and topographic properties of single neurons in cats somatic sensory cortex. Journal of Neurophysiology 20:408–434
Mountcastle V (1997) The columnar organization of the neocortex. Brain 120:701–722
Nauta WJ, Karten HJ (1970) A general profile of the vertebrate brain, with sidelights on the ancestry of cerebral cortex. In: Schmitt FO (ed) The Neurosciences: Second Study Program, Rockfeller University Press, New York, pp 1–7
Nelson RJ (2002a) The somatosensory system: deciphering the brain's own body image. CRC Press, Boca Raton (FL)
Nelson SB (2002b) Cortical microcircuits: Diverse or canonical? Neuron 36:19–27
Newcomb RW (1994) Some historical perspectives on early pulse coded neural network circuits. In: Zaghloul ME, Meador JL, Newcomb RW (eds) Silicon implementation of pulse coded neural networks, Springer-Verlag, Berlin, pp 1–8
Nieder A (2009) Prefrontal cortex and the evolution of symbolic reference. Current Opinion in Neurobiology 19:99–108
Noack RA (2012) Solving the "human problem": The frontal feedback model. Consciousness and Cognition 21:1043–1067

Papineau D (1993) Philsophical Naturalism. Basil Blackwell, Oxford (UK)
Paynter H, Beaman JJ (1991) On the fall and rise of the circuit concept. Journal of the Franklin Institute 328:525–534
Paz AW (2017) A mechanistic perspective on canonical neural computation. Philosophical Psychology 30:209–230
Peters A, Payne BR (1993) Numerical relationships between geniculocortical afferents and pyramidal cell modules in cat primary visual cortex. Cerebral Cortex 64:467–478
Pettersen KH, Einevoll GT (2009) Neurophysics: what the telegrapher's equation has taught us about the brain. In: Martinsen OG, Jensen O (eds) An anthology of developments in clinical engineering and bioimpedance – Festschrift for Sverre Grimnes, Medisin-teknisk avdelings forlag, Oslo, pp 94–108
Phillips C, Powell T, Shepherd GM (1963) Responses of mitral cells to stiulation of the lateral olfactory tract in the rabbit. Journal of Physiology 168:65–88
Piccinini G (2006) Computational explanation in neuroscience. Synthese 153:343–353
Piccinini G (2007) Computational modeling vs. computational explanation: Is everything a Turing Machine, and does it matter to the philosophy of mind? Australasian Journal of Philosoph 85:93–115
Piccinini G (2015) Physical Computation: A Mechanistic Account. Oxford University Press, Oxford (UK)
Plebe A (2012) A model of the response of visual area V2 to combinations of orientations. Network: Computation in Neural Systems 23:105–122
Potjans TC, Diesmann M (2014) The cell-type specific cortical microcircuit: Relating structure and activity in a full-scale spiking network model. Cerebral Cortex 24:785–806
Priebe NJ, Ferster D (2006) Mechanisms underlying cross-orientation suppression in cat visual cortex. Nature Neuroscience 9:552–562
Prinz J (2006) Is the mind really modular? In: Stainton RJ (ed) Contemporary Debates In Cognitive Science, Blackwell Publishing, Malden (MA), pp 22–36
Pulvermüller F (2002) The Neuroscience of Language: On Brain Circuits of Words and Serial Order. Cambridge University Press, Cambridge (UK)
Pylyshyn Z (1981) Computation and cognition: Issues in the foundations of cognitive science. Behavioral and Brain Science 3:111–150
Rakic P (1971) Guidance of neurons migrating to the fetal monkey neocortex. Brain Research 33:471–476
Rakic P (1995) Radial versus tangential migration of neuronal clones in the developing cerebral cortex. Proceedings of the Natural Academy of Science USA 92:323–327
Rakic P (2008) Confusing cortical columns. Proceedings of the Natural Academy of Science USA 34:12,099–12,100
Rall W (1957) Membrane time constant of motoneurons. Science 126:454
Rall W (1969) Time constants and electrotonic length of membrane cylinders and neurons. Biophysic Journal 9:1483–1508

Rall W (2006) Wilfrid rall. In: Squire LR (ed) The History of Neuroscience in Autobiography, Elsevier, Amsterdam, pp 552–661, volume 5

Rall W, Shepherd GM (1968) Theoretical reconstruction of field potentials and dendrodendritic synaptic interactions in olfactory bulb. Journal of Neurophysiology 31:884–915

Ramón y Cajal S (1891) On the structure of the cerebral cortex in certain mammals. La Cellule 7:125–176

Rathkopf CA (2013) Localization and intrinsic function. Philosophy of Science 80:1–21

Rauschecker JP, Tian B, Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268:111–114

Reynolds JH, Chelazzi L, Desimone R (1999) Competitive mechanisms subserve attention in macaque areas V2 and V4. Nature Neuroscience 2:1019–1025

Rockel A, Hiorns R, Powell T (1974) Numbers of neurons through full depth of neocortex. Journal of Anatomy 118:371

Rockel A, Hiorns R, Powell T (1980) The basic uniformity in structure of the neocortex. Brain 103:221–244

Rockland KS (2011) Five points on columns. Frontiers in Neuroanatomy 4:Article 22

Roe AW, Garraghty P, Sur M (1987) Retinotectal W cell plasticity: experimentally induced retinal projections to auditory thalamus in ferrets. Soc Neurosci Abst 13:1023

Roe AW, Garraghty P, Esguerra M, Sur M (1990) A map of visual space induced in primary auditory cortex. Science 250:818–820

Rose D (1979) Mechanisms underlying the receptive field properties of neurons in cat visual cortex. Vision Research 19:533–544

Rose N, Abi-Rached JM (2013) Neuro: The New Brain Sciences and the Management of the Mind. Princeton University Press, Princeton (NJ)

Rothschild G, Mizrahi A (2015) Global order and local disorder in brain maps. Annual Review of Neuroscience 38:247–268

Roux E (2014) The concept of function in modern physiology. Journal of Physiology 592:2245–2249

Shepherd GM (1974) The Synaptic Organization of the Brain. Oxford University Press, Oxford (UK)

Shepherd GM (1979) The Synaptic Organization of the Brain. Oxford University Press, Oxford (UK), second Edition

Shepherd GM (1988) A basic circuit for cortical organization. In: Gazzaniga MS (ed) Perspectives on Memory Research, MIT Press, Cambridge (MA), pp 93–134

Shepherd GM (1991) Foundations of the neuron doctrine. Oxford University Press, Oxford (UK)

Shepherd GM (2004a) Introduction to synaptic circuits. In: Shepherd (2004b), pp 1–38, 5th Edition

Shepherd GM (ed) (2004b) The Synaptic Organization of the Brain. Oxford University Press, Oxford (UK), 5th Edition

Shepherd GM (2016) Foundations of the neuron doctrine. Oxford University Press, Oxford (UK), 25th anniversary edition

Smith CU (1992) A century of cortical architectonics. Journal of the History of the Neurosciences 1:201–218

Sober E (1999) The multiple realizability argument against reductionism. Philosophy of Science 64:542–564

Song S, Abbott LF (2001) Cortical development and remapping through spike timing-dependent plasticity. Neuron 32:339–350

Sur M (1989) Visual plasticity in the auditory pathway: Visual inputs induced into auditory thalamus and cortex illustrate principles of adaptive organization in sensory systems. In: Dynamic Interactions in Neural Networks: Models and Data, Springer-Verlag, Berlin, pp 35–52

Thomson AM, West DC, Wang Y, Bannister P (2002) Synaptic connections and small circuits involving excitatory and inhibitory neurons in layers 2-5 of adult rat and cat neocortex: Triple intracellular recordings and biocytin labelling *in vitro*. Cerebral Cortex 12:936–953

Thomson Kelvin W (1842) On the uniform motion of heat in homogeneous solid bodies and its connection with the mathematical theory of electricity. Cambridge mathematical journal 3:71–84

Thomson Kelvin W (1855) On the theory of the electric telegraph. Proceedings of the Royal Society of London 7:382–399

Turing A (1936) On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society 42:230–265

Verplaetse J, Schrijver JD, Vanneste S, Braeckman J (eds) (2009) The Moral Brain Essays on the Evolutionary and Neuroscientific Aspects of Morality. Springer-Verlag, Berlin

Vogt C, Vogt O (1903) Zur anatomische Gliederung des Cortex Cerebri. Journal Psychol Neurol 2:160–180

Vogt C, Vogt O (1919) Allgemeine Ergebnisse unserer Hirnforschung. Journal Psychol Neurol 25:279–461

von der Malsburg C (1973) Self-organization of orientation sensitive cells in the striate cortex. Kybernetic 14:85–100

von Economo C, Koskinas GN (1925) Die Cytoarchitektonik der Hirnrinde des erwachsenen Menschen. Springer-Verlag, Berlin

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Willshaw D (2006) Self-organization in the nervous system. In: Morris R, Tarassenko L (eds) Cognitive Systems: Information Processing Meets Brain Science, Elsevier, Amsterdam, pp 5–33

Willshaw DJ (1972) A simple network capable of inductive generalization. Proceedings of the Royal Society of London B 182:233–247

Willshaw DJ, Hallam J, Gingell S, Lau SL (1997) Marr's theory of the neocortex as a self-organizing neural network. Neural Computation 9:911–936

Wilson MA, Bower JM (1989) The simulation of large-scale neural networks. In: Koch C, Segev I (eds) Methods in Neuronal Modeling, MIT Press, Cambridge (MA), pp 291–333

Wright JJ, Bourke PD (2017) Further work on the shaping of cortical development and function by synchrony and metabolic competition. Frontiers in Computational Neuroscience 10:Article 127

Wright L (1976) Teleological Explanations. University of California Press, Berkeley (CA)

Young MP, Hilgetag CC, Scannell JW (2000) On imputing function to structure from the behavioural effects of brain lesions. Philosophical transactions of the Royal Society B 355:147–161