Nat Comput (2012) 11:141–150
DOI 10.1007/s11047-012-9306-0

On reverse engineering in the cognitive and brain sciences

Andreas Schierwagen

Published online: 10 February 2012

Abstract Various research initiatives try to utilize the operational principles of organisms and brains to develop alternative, biologically inspired computing paradigms and artificial cognitive systems. This article reviews key features of the standard method applied to complexity in the cognitive and brain sciences, i.e. decompositional analysis or reverse engineering. The indisputable complexity of brain and mind raise the issue of whether they can be understood by applying the standard method. Actually, recent findings in the experimental and theoretical fields, question central assumptions and hypotheses made for reverse engineering. Using the modeling relation as analyzed by Robert Rosen, the scientific analysis method itself is made a subject of discussion. It is concluded that the fundamental assumption of cognitive science, i.e. complex cognitive systems can be analyzed, understood and duplicated by reverse engineering, must be abandoned. Implications for investigations of organisms and behavior as well as for engineering artificial cognitive systems are discussed.

Keywords Brain · Cognition · Capacity · Decompositional analysis · Localization · Linearity · Modularization · Column concept · Reverse engineering · Complex systems · Modeling relation

A. Schierwagen (✉)
Intelligent Systems Department, Institute for Computer Science, University of Leipzig, Leipzig, Germany
e-mail: schierwa@informatik.uni-leipzig.de

1 Introduction

For some time past, computer science and engineering devote close attention to the functioning of the brain. It has been argued that recent advances in cognitive science and neuroscience have enabled a rich scientific understanding of how cognition works in the human brain. Thus, research programs have been initiated by leading research organizations to build new computing systems based on information processing principles derived from the working of the brain, and to develop new cognitive architectures and computational models of human cognition (see, e.g. (Schierwagen 2007, 2009), and references therein).

Two points are emphasized in those research programs: First, there is impressive abundance of available experimental brain data, and second, we have the computing power to meet the enormous requirements to simulate a complex system like the brain. Given the improved scientific understanding of the operational principles of the brain as a complexly organized system, it should then be possible to build an operational, quantitative model of the brain. Tuning the model could be achieved then using the deluge of empirical data, due to the ever-improving experimental techniques of neuroscience.

Trying to put this idea into practice, however, has generally produced disenchantment after high initial hopes and hype. If we rhetorically pose the question "What is going wrong?" [as previously posed in the field of robotics (Brooks 2001)], possible answers are: (1) The parameters of our models are wrong; (2) We are below some complexity threshold; (3) We lack computing power; (4) We are missing something fundamental and unimagined. In most cases, only answers (1–3) are considered by computer and AI scientists, and allied neuroscientists, and conclusions are drawn in similar vein. If answer (1) is considered

true, still better experimental methodologies are demanded to gather the right data, preferably at the molecular genetic level [e.g. (Le Novere 2007)]. Answers (2) and (3) often induce claims for concerted, intensified efforts relating phenomena and data at many levels of brain organization [e.g. (Grillner et al. 2005)].

Together, any of answers (1–3) would mean that there is nothing *in principle* that we do not understand about brain organization. All the concepts and components are present, and need only to be put into the model. This view is widely taken; it represents the belief in the efficiency of the scientific method, and it leads one to assume that our understanding of the brain will major advance as soon as the 'obstacles' are cleared away.

As I will show in this article, there is, however, substantial evidence in favour of answer (4). I will argue that, by following the standard scientific method, we are in fact ignoring something fundamental, namely that biological and engineered systems are basically different in nature.

The article is organized as follows. Section 2 presents conceptual and methodological basics of the cognitive and brain sciences. The concepts of decompositional analysis and localization underlying the reverse engineering method are reviewed. I discuss the idea of modularization and its relation to the superposition principle of system theory. Then, Sect. 3 shortly touches on Blue Brain and SyNAPSE, two leading reverse-engineering projects. Both projects are based on the hypothesis of the columnar organization of the cortex. The rationale underlying reverse engineering in cognitive and brain sciences is outlined. New findings are mentioned questioning the concept of the basic uniformity of the cortex, and consequences for the reverse-engineering projects are discussed. Section 4 ponders about the claim that non-decomposability is not an intrinsic property of complex systems but is only in our eyes, due to insufficient mathematical techniques. For this, the modeling relation as analyzed by Robert Rosen is explained which enables us to make the scientific analysis method itself a subject of discussion. It is concluded that the fundamental assumption of cognitive science must be abandoned. I end the article by some conclusions for the study of organisms and behavior as well as for engineering artificial cognitive systems.

2 Methodological basics

2.1 Decomposability

Brains, even those of simple animals, are enormously complex structures, and it is a very ambitious goal to cope with this complexity. The scientific disciplines involved in cognitive and brain research are committed to a common methodology to explain the properties and capacities of complex systems. It is decompositional analysis, i.e. analysis of the system in terms of its components or subsystems.

Since Simon's influential book "The Sciences of the Artificial" (Simon 1969), (near-) decomposability of complex systems has been accepted as fundamental for the cognitive and brain sciences. Cognitive capacities are considered as dispositional properties which can be explained via decompositional analysis. I call this the *fundamental assumption* for the cognitive and brain sciences. Simon (1969), Wimsatt (1986) and Bechtel and Richardson (1993), among others, have further elaborated this concept. They consider decomposability a continously varying system property, and state, roughly, that systems fall on a continuum from aggregate (full decomposable) to integrated (non-decomposable) (Fig. 1). The *fundamental assumption* implies that cognitive and brain systems are non-ideal aggregate systems; the capacities of the components are internally realized by strong intra-component interactions, and interactions between components do not appreciably contribute to the capacities; they are much weaker than the intra-component interactions. Hence, the description of the complex system as a set of weakly interacting components seems to be a good approximation. This property of complex systems, which should have evolved through natural selection, was called near-decomposability (Simon 1969).

Simon characterizes near-decomposability as follows: (1) In a nearly decomposable system, the short-run behaviour of each of the component subsystems is approximately independent of the short-run behaviour of the other components; (2) in the long run the behaviour of any one of the components depends in only an aggregate way on the behaviour of the other components (Simon 1969, p. 100). Thus, if the capacities of a near-decomposable system are to be explained, to some approximation its components can be studied in isolation, and based on their known interactions, their capacities eventually combined to generate the systems behavior.

Let us summarize this assumption because it is of central importance in the following:

Fundamental assumption for cognitive and brain sciences

Cognitive and brain systems are non-ideal aggregate systems. The capacities of the components are

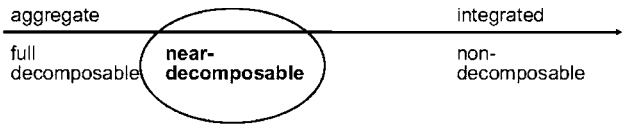

Fig. 1 Decomposability as a continuously varying system property. According to this view, the focus is on near-decomposable systems which would represent the most relevant systems category in the cognitive and brain sciences

internally realized (strong intra-component interactions) while interactions between components are negligible with respect to capacities. Any capacity of the whole system then results from superposition of the capacities of its subsystems. This property of cognitive and brain systems should have evolved through natural selection and is called near-decomposability.

2.2 Decompositional analysis

The primary goal of cognitive science and its subdisciplines is to understand cognitive capacities like vision, language, memory, planning etc. That is, we want to answer questions of the form "does system S possess or exercise a capacity C?". The quest for S's capacity C can be replaced by evaluating the proposition $P(S)$ = "S possesses or exercises the capacity C". In other words, we want to determine the truth or falsity of the proposition $P(S)$.

Cummins (1983, 2000) suggests that a system's capacity can be explained by a *functional analysis* of that capacity. A functional analysis of some capacity C proceeds, roughly, by parsing the capacity into a set of constituent sub-capacities $C_1, C_2, \ldots, C_n$. Note that the sequence has to be specified in which those constituent capacities must be exercised for producing the complex capacity. That is, there is an algorithm which can be programmed to decide whether system S has C or P, by processing a finite list of propositions $P_1, P_2, \ldots, P_n$.

The scheme then asserts that any capacity proposition $P(S)$ can be expressed as conjunction of a finite number of sub-propositions $P_i(S)$ the truth of each one is necessary, and all together sufficient, for $P(S)$ to be true[1]. Hence, a functional analysis comprises the following steps:

Functional analysis

1. Establish that system S has capacity C or property P.
2. Decompose P into sub-properties $P_1(S), P_2(S), \ldots, P_n(S)$.
3. Specify the sequence in which the sub-properties P_i are to be processed to generate P, i.e. the algorithm.

Then it follows that $P(S) = \bigwedge_{i=1}^{n} P_i(S)$. (1)

If this scheme is applied to a material system S with the property $P(S)$, it allows to express $P(S)$ in the form of Eq. 1, i.e. by purely syntactical means. That is, property $P(S)$ is redundant, and its truth does not provide new information about system S, cf. (Rosen 2000).

[1] Cummin's scheme evidently employs Frege's *principle of compositionality*, well-known in computer science as 'divide and conquer'.

A cognitive capacity may be explained not only by analyzing the capacity itself, but also by analyzing the system that has it. This type of decompositional analysis is *structural analysis* (Atkinson 1998; Eckardt 2004). It involves to attempt to identify the structural, material components of the system. Thus, the material system S is to be decomposed into *context-independent* components S_j, i.e. their individual properties $P_k(S_j)$ are independent of the decomposition process itself and of S's environment.

2.3 Localization

Functional analysis and structural analysis must be clearly differentiated, although in practice, there is a close interplay between them (as indicated by the double arrow in Fig. 2). This is obvious in the *localization approach* which combines both analysis types. The essential assumption is that each of the sub-properties $P_i(S)$, into which the property $P(S)$ was decomposed, is to be localized in some particular subsystem S_j of S itself. Thus, the properties $P_k(S_j)$ of S's material components S_j equal exactly the conjunction terms in Eq. 1, i.e. for each sub-property $P_i(S)$ there is a structural component S_j whose property $P_k(S_j)$ is identical to $P_i(S)$,

$$P_i(S) = P_k(S_j). \quad (2)$$

Thus, Eq. 1 can be rewritten as

$$P(S) = \bigwedge_{j,k} P_k(S_j). \quad (3)$$

Equation 2 in a nutshell expresses the idea of the *fundamental assumption*, i.e. decomposition and localization. Properties $P_i(S)$ of the whole system S are identified with properties $P_k(S_j)$ of certain of its subsystems S_j. This means, one assumes that any material system S (including brain) can be decomposed into context-independent parts or structural

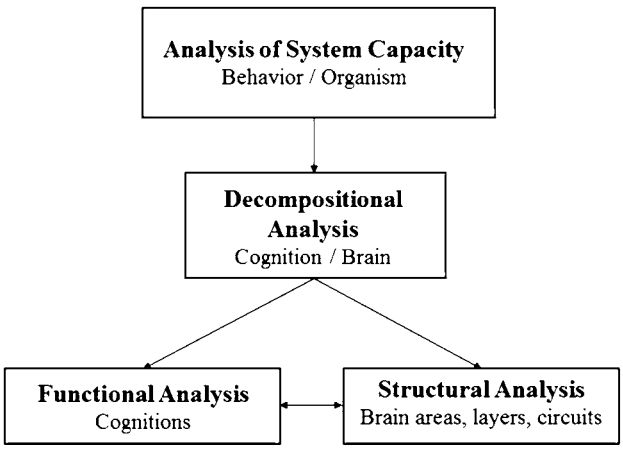

Fig. 2 View on decompositional analysis of brain and cognition. See text for details

components S_j in such a way that their properties $P_k(S_j)$ are independent of the properties of the other parts and of any environment. Thus, a set of decomposition operators $\mathscr{D}_i$ on S of the form

$$\mathscr{D}_i(S) = S_i \tag{4}$$

is supposed which isolate the subsystems S_i from S. Corresponding to the *fundamental assumption*, the operators $\mathscr{D}_i$ break the inter-component interactions which 'glue' the context-independent components of S together, but without affecting any of the intra-component interactions.

Ideally, decomposition operations like $\mathscr{D}_i$ are reversible, i.e. the whole system S can be synthesized from the components S_j,

$$S = S_1 \otimes S_2 \otimes \cdots \otimes S_m, \tag{5}$$

where the $\otimes$-symbol denotes inter-component interactions like those broken by the decomposition operators $\mathscr{D}_i$. Thus, S is to be considered as a kind of direct product. Now the close analogy of expressions (3) and (5) becomes obvious: the fractionation of system S corresponds to the compositionality of property $P(S)$ while the connector symbol $\wedge$ replaces the inter-component interaction symbol $\otimes$,

$$P(S) = P_1(S_1) \wedge P_2(S_2) \wedge \cdots \wedge P_m(S_m). \tag{6}$$

These suppositions allow to proceed wholly in the syntactical realm. Any property P of a physical system S comes with an algorithm for evaluating P's truth, and any physical system S can be algorithmically generated from a sufficiently large population of components S_i by exclusively syntactical means. In both cases, analysis and synthesis are inverse operations which are realized entirely by algorithms, i.e. the operations are computable, cf. (Rosen 2000, p. 131).

Understating the case, the localization approach has been described as hypothetical identification which is to serve as research heuristics (Bechtel 1993). In fact, however, the majority of cognitive scientists considers it as fundamental and indispensable [e.g. (Eckardt 2004; Ross 2010)]. For example, Von Eckardt (2004) points out that a functional analysis for a capacity C only provides us with a *possible* explanation of how the system has capacity C. That is because the decomposition of a composed capacity is not unique—it can be parsed into various alternative sequences of constituent capacities, each of which is sufficient for S's capacity C. As a way out, she suggests to build a model that is structurally adequate by employing the localization approach.

A caveat is necessary, however: There is no reason to assume that functional and structural components match up one-to-one! Of course, it might be the case that some functional components map properly onto individual structural components. It is rather probable, however, for a certain functional component to be implemented by non-localized, spatially distributed material components. Conversely, a given structural component may implement more than one distinct function. According to Dennett (1991, p. 273): "In a system as complex as the brain, there is likely to be much 'multiple, superimposed functionality'". With other words, we cannot expect specific functions to be mapped to structurally bounded neuronal structures, and vice versa. It is now well known that Dennett's caveat has been proved as justified (e.g. (Price 2005)). Thus, the value of the localization approach as 'research heuristics' seems rather dubious (Uttal 2001; Henson 2005).

2.4 Linearity, modularization and complex systems

In the cognitive and brain sciences, as in other fields of science, the components of near-decomposable systems are called modules. This term originates from engineering; it denotes the process of decomposing a product into building blocks, modules, with specified interfaces, driven by the designer's interests and intended functions of the product. It refers either to functional or structural components. Modularized systems are linear in the sense that they obey an analog of the superposition principle of linear system theory in engineering (Schierwagen 1989). If the modules are structurally localized functional components, the superposition principle is expressed by Eq. 5. The function of a decomposable system results from the linear combination of the functions of the system modules[2] This principle mirrors the constructive step in the scheme of reverse engineering (see above and Sect. 3 below). The terms 'linear' and 'nonlinear' are often used in this way: 'Linear' systems are decomposable into independent modules with linear, proportional interactions while 'nonlinear' systems are not[3] (Schierwagen 1989; Forrest 1990).

Applying this concept to the systems at the other end of the complexity scale (Fig. 1), the integrated systems are basically not decomposable, due to the strong, nonlinear interactions involved. Thus, past or present states or actions of any or most subsystems always affect the state or action of any or most other subsystems. In practice, analyses of integrated systems nevertheless try to apply the methodology for decomposable systems, in particular if there is some hope that the interactions can be linearized. Such linearizable systems have been above denoted as nearly decomposable. However, in the case of strong nonlinear

[2] A corresponding class of models in mathematics is characterized by the superposition theorem for homogeneous linear differential equations stating that the sum of any two solutions is itself a solution.

[3] We must differentiate between the natural, complex system and its description using modeling techniques from linear system theory or nonlinear mathematics.

interactions, we must accept that decompositional analysis is not applicable.

Already several decades ago this insight was stressed. For example, Levins (1970, p. 76 ff.) around 1970 proposed a classification of systems into aggregate, composed and evolved systems. While the aggregate and the composed would not cause serious problems for decompositional analyses, Levins emphasized the special character of evolved systems:

> A third kind of system no longer permits this kind of analysis. This is a system in which the component subsystems have evolved together, and are not even obviously separable; in which it may be conceptually difficult to decide what are the really relevant component subsystems.... The decomposition of a complex system into subsystems can be done in many ways... it is no longer obvious what the proper subsystems are, but these may be processes, or physical subsets, or entities of a different kind.

This statement clearly contradicts the *fundamental assumption*, and it has not lost its relevance, as the findings of complexity science have shown. Nevertheless, most researchers in the cognitive and brain sciences found reasons to cling to it. A main argument for the *fundamental assumption* has been that non-decomposability is only in our eyes, and not an intrinsic property of strongly nonlinear systems, and scientific progress will provide us with the new mathematical techniques required to deal with nonlinear, integrated systems. I will return to this problem in Sect. 4.

3 Reverse engineering the brain

3.1 The column concept

A guiding idea about the composition of the brain is the hypothesis of the columnar organization of the cerebral cortex. This *column concept* was developed mainly by Hubel and Wiesel (1963), Mountcastle (1997) and Szenthágothai (1983), and later on, it was published in the influential paper "The basic uniformity in structure of the neocortex" (Rockel et al. 1980). According to this hypothesis (which has been taken more or less as fact by many experimental as well as theoretical neuroscientists), the neocortex is composed of 'building blocks' (Fig. 3) of repetitive structures, the 'columns' or 'canonical cortical circuits', and it is characterized by a basic canonical pattern of connectivity. In this scheme all cortical areas would perform identical or similar computational operations with their inputs.

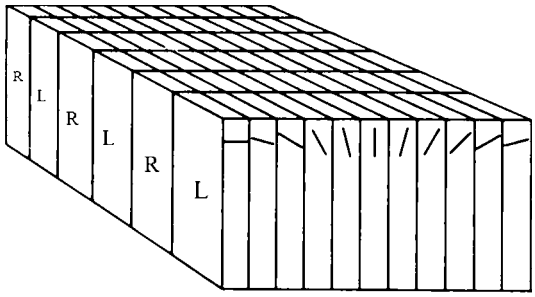

Fig. 3 Hubel and Wiesel's 'ice cube' model of visual cortical processing. The diagram illustrates the idea that the cortex is composed of iterated modules each of which comprises a complete set of superimposed feature-processing elements, in this case for ocular dominance (indicated by L and R) and orientation selectivity (here represented for angles from 0 to π) (after (Hubel 1977))

3.2 Method of reverse engineering

Referring to and based on these works, several projects started recently, among them the *Blue Brain Project* (Markram 2006) and the *SyNAPSE Project* (Systems of Neuromorphic Adaptive Plastic Scalable Electronics, SyNAPSE). They are considered to be "attempts to reverse-engineer the mammalian brain, in order to understand brain function and dysfunction through detailed simulations" (Markram 2006) or, more pompous, "to engineer the mind" (SyNAPSE 2008).

Reverse engineering is the main method used in empirical research to integrate the data derived from the different levels of the brain organization. Originally a concept in engineering and computer science, reverse engineering involves as first step a *decompositional analysis*, i.e. the detailed examination of a functional system (*functional analysis*) and its dissecting at the physical level into component parts (*structural analysis*), see Fig. 2. In a second step, the (re-) construction of the original system is attempted by creating duplicates including computer models, see below (Sect. 4). This method is usually not much discussed with respect to its assumptions, conditions and range[4] but see (Dennett 1994; Marom et al. 2009; Gurney 2009).

The central role in these projects play cortical microcircuits or columns. As Maas and Markram (2006) formulate, it is a "tempting hypothesis regarding the computational role of cortical microcircuits ... that there exist genetically programmed stereotypical microcircuits that compute certain basis function." Their study well

[4] Only recently, differences between proponents of reverse engineering on how it is appropriately to be accomplished became public. The heads of the two reverse engineering projects mentioned, Markram (2006) and Modha (Systems of Neuromorphic Adaptive Plastic Scalable Electronics, SyNAPSE 2008), disputed publicly as to what granularity of the modeling is needed to reach a valid simulation of the brain. Markram questioned the authenticity of Modha's respective claims (Brodkin 2009).

illustrates the modular approach fostered, e.g. by (Grillner et al. 2005; Gurney 2009; Arbib et al. 1997; Bressler 2006). Invoking the localization concept, the tenet is that there exist fundamental correspondences among the anatomical structure of neuronal networks, their functions, and the dynamic patterning of their active states. Starting point is the 'uniform cortex' with the cortical microcircuit or column as the structural component. The question for the functional component is answered by assuming that there is a one-to-one relationship between the structural and the functional component (see Sect. 2.2). Together, the modularity hypothesis of the brain is considered to be both structurally and functionally well justified.

As quoted above, the goal is to examine the hypothesis that there exist genetically programmed stereotypical microcircuits that compute certain basis function, thus providing for complex cognitive capacities. This hypothesis is based on the general, computational approach to cognitive capacities which takes for granted that "cognition is computation", i.e. the brain produces the cognitive capacities by computing functions[5].

This assumption allows to apply the idea of decomposition or reverse engineering in the following way. From mathematical analysis and approximation theory it is well-known that a broad class of practically relevant functions f can be approximated by composition or superposition of some basis functions. Of prime relevance in this respect are Kolmogorov's 'superposition theorem' stating that continuous functions of n arguments can always be represented using a finite composition of functions of a single argument, and addition, and Weierstrass and Stone's classical result that any real continuous function can be approximated with arbitrary precision using a finite number of computing units. Kolmogorov's theorem was rediscovered in the 1980s by Hecht-Nielsen and applied to artificial neural networks. Since then many different types of *function networks* and their properties have been investigated, so that a lot of results about the approximation properties of networks are already available, see e.g. (Suykens et al. 1996).

For example, neural networks for function approximation have been developed based on orthogonal basis functions such as Fourier series, Bessel functions and Legendre polynomials. A typical configuration of such neural network of feedforward type is illustrated in Fig. 4. The input is $\bar{x} = (x_1, \ldots, (x_m)$ and the output is $f(\bar{x}) = \bar{y} = (y_1, \ldots, y_k)$ with

$$y_j = w_{1,j} \cdot f_1(\bar{x}) + w_{2,j} \cdot f_2(\bar{x}) + \cdots + w_{n,j} \cdot f_n(\bar{x}) \qquad (7)$$

The functions $f_i(\bar{x})(i = 1, \ldots, n)$ are the basis functions computed by the network units. The real numbers $w_{i,j}$ $(i = 1, \ldots, n; j = 1, \ldots, k)$ are their respective weights which can be adapted using effective learning algorithms to approximate the function $f(\bar{x})$.

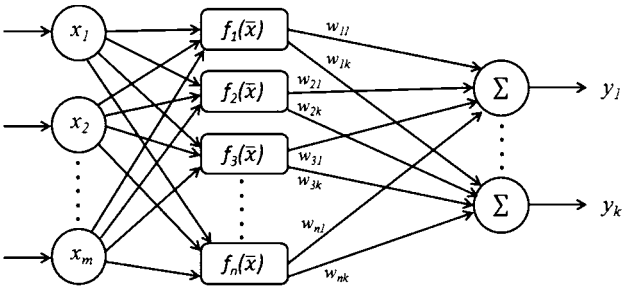

Fig. 4 Example of a network computing the function $f(\bar{x}) = \bar{y} = (y_1, \ldots, y_k)$

As one can see, Eq. 7 represents the analog of Eqs. 3 and 5, now in the computational realm. That is, from functional analysis and decomposition of a cognitive capacity into subcapacities, and fractionation of the cortex (or some subsystem) into parts we arrive at linear decomposition of a cognitive function into elements of a given set of 'simple', or basis functions.

Thus, if some basis functions were identified, they provided the components of a (possible) computational decomposition. The reverse engineering method as applied in the cognitive and brain sciences from a computational perspective then proceeds as follows:

Reverse engineering the cortex

1. Capacity analysis: Specify a certain cognitive capacity which is assumed to be produced through the cortex by computing a certain function.
2. Decompositional analysis:
   (a) Functional (computational) analysis: Select a set of basis functions which might serve as functional components or computational units in the cortex.
   (b) Structural analysis: Identify a set of anatomical components of the cortex. Provide evidence that cortical microcircuits are the anatomical components of the cortex.
3. Localization: Provide evidence for the functional components or computational units being linked with the anatomical components.
4. Synthesis:
   (a) Modeling:
      i. Establish a structurally adequate functional model of the computational unit (the presumed 'canonical circuit') which generates the basis functions specified in step 2.(a).

[5] See (Schierwagen 2007) for discussion of the computational approaches (including the neurocomputational one) to brain function, and their shortcomings.

ii. Build a structurally adequate network model of the cortex (or some subsystem) composed of the canonical circuit models.

(b) Simulation: Prove that the specific cognitive capacity or function under study is generated by the network of circuit models, i.e through superposition of the specified basis functions.

3.3 Hypotheses and reality

With the reverse engineering scheme formulated above, we have a 'recipe' at hand which could facilitate the analysis very much. Recent findings in the experimental and theoretical fields, however, have objected most of the assumptions and hypotheses made as problematic, if not inappropriate and unrealistic. Already step 1, specification of a cognitive capacity, poses serious problems. It has always been extremely difficult to define exactly what is meant by a psychological, cognitive, or mental term, and the possibility should be acknowledged that they are only figments of our experimental designs or convenient artifices to organize our theoretical models (Uttal 2001). This difficulty is obvious in recent attempts to build *cognitive ontologies* (e.g. (Price 2005; Henson 2005).

Likewise, the assumptions about the structural and functional composition of the cortex, the notion of the basic uniformity in the cortex with respect to the density and types of neurons per column for all species turned out to be untenable (e.g. Horton 2005; Rakic 2008; Herculano-Housel et al. 2008; Frégnac 2006). It has been impossible to find the cortical microcircuit that computes a specific basis function (Frégnac 2006; de Garis et al. 2010). No genetic mechanism has been deciphered that designates how to construct a column. The column structures encountered in many species (but not in all) seem to represent spandrels (structures that arise non-adaptively, i.e. as an epiphenomenon) in various stages of evolution (Gould 1979).

Step 4-synthesis—is worth extended discussion which space limitations forbid. In short, this step represents the conviction that large-scale modeling of brain networks will eventually lead to understanding the mind-brain problem. It has been argued that producing and understanding complex phenomena from the interaction of simple nonlinear elements like artificial neurons or cellular automata is possible. One expects then, that this would also work for cortical circuits which are recognized as nonlinear devices, and theories could be applied (or developed, if not yet available) that would guide us to which model setup might have generated a given network behavior. This would complete the reverse engineering process.

However, findings in complexity or nonlinear science exclude such transfer of the usual, linear approach. It is now clear that finding out which processes caused a specific complex behavior of a given system—an *inverse problem*—is hard because of its *ill-posedness*[6]. This means for the study of cortical circuits and networks of them that from observed activity or function we cannot, in principle, infer the internal organization. A wide variety of different organizations can produce the same behavior Edmonds (2009).

If we revisit the column concept of the cortex employed in theories of brain organization, we recognize that hypothesized structural components (cortical columns) have been identified with alike hypothetical functional components (basis function), employing the localization concept (Sect. 2.2). As we have seen, the facts contradict these assumptions, i.e. the reverse engineering project has been based on false presuppositions. In contrast to the localization idea, there is evidence for a given functional component to be implemented by spatially distributed networks and, vice versa, for a given structural component to implement more than one distinct function. With other words, it is not feasible for specific functions to be mapped to structurally bounded neuronal structures (Price 2005; Horton 2005; Rakic 2008; Herculano-Housel et al. 2008).

This means, although the column concept is an attractive idea both from neurobiological and computational point of view, it cannot be used as an unifying principle for understanding cortical function. Thus, it has been concluded that the concept of the cortex as a 'large network of identical units' should be replaced with the idea that the cortex consists of 'large networks of diverse elements' whose cellular and synaptic diversity is important for computation (e.g. (Frégnac 2006).

It is worth to notice that the claim for conceptual change towards 'cortex as large network of diverse elements' completely remains within the framework of reverse engineering, i.e. it is a plea for 'Just carry on!'. It appears questionable, however, that the original goals of the cognitive and brain sciences and AI can be achieved this way. Actually, the methods of decompositional analysis and reverse engineering themselves have been principally criticized, which will be shortly discussed in the next section.

4 The modeling relation

In Sect. 2.4, I concluded that complex, integrated systems are basically non-decomposable, thus resisting the standard analysis method. Now I return to this issue and to the

[6] In mathematics, a problem is called ill-posed if no solution or more than one solution exists, or if the solutions depend discontinuously upon the initial data.

consequences for investigating such systems in the cognitive and brain sciences.

Despite contradicting findings in complex systems science, the majority of researchers in the cognitive and brain sciences subscribes for the *fundamental assumption*, i.e. the relevant systems in the cognitive and brain sciences are treated as nearly decomposable. Accordingly, non-decomposability is considered not as intrinsic property of complex, integrated systems but only as subjective, temporary failure of our methodology, due to insufficient mathematical techniques (e.g. Bechtel 2002).

In contrast to that, Rosen (1991; Rosen 2000) has argued that understanding complex, integrated systems requires making the scientific analysis method itself a subject of discussion. A powerful method of understanding and exploring the nature of the scientific method, and in particular, reverse engineering, provides the *modeling relation*. It is this relation by which scientists bring "entailment structures into congruence" (Rosen 1991, p. 152). The modeling relation is represented by the set of mappings shown in Fig. 5. It relates two systems, a natural system N and a formal system F, by a set of arrows depicting processes or mappings. The assumption is that this diagram represents the various processes which we are carrying out when we perceive the world.

The modeling relation is a relation in the formal mathematical sense,

$$R = \{(a, c) \mid a = b \circ c \circ d\} \tag{8}$$

while $\circ$ denotes concatenation. The members a and c of each ordered pair in R are entailments from the two systems, N and F. Natural system N is part of the physical world that we wish to understand (in our case: human being, organism, brain), in which things happen according to rules of causality (arrow a). That is, if some cause acts on N, then the system will behave in a certain way, or produce certain effects. This resultant coupling of cause and effect in N is called *causal entailment*.

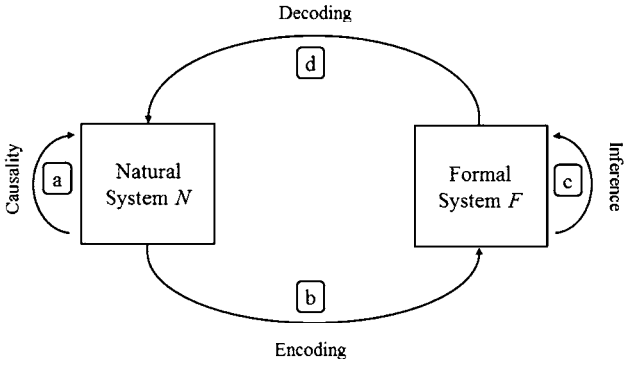

Fig. 5 The modeling relation. A natural system N is modeled by a formal system F. Each system has its own internal entailment structures (*arrows a* and *c*), and the two systems are connected by the encoding and decoding processes (*arrows b* and *d*). After (Rosen 1991, p. 60)

On the right of Fig. 5, F represents symbolically the parts of the natural system (observables) which we are interested in, along with formal rules of inference (arrow c) that essentially constitute our working hypotheses about the way things work in N, i.e. the way in which we manipulate the formal system to try to mimic causal events observed or hypothesized in the natural system on the left. Stated another way, F has inferential linkage; that is, if some premise proposition acts on F, then it will generate a consequential proposition as conclusion. This resultant coupling of premise and conclusion in F is called *inferential entailment*.

Arrow b represents the encoding of the parts of N under study into the formal system F, i.e. a mapping that establishes the correspondence between observables of N and symbols defined in F. Predictions about the behavior in F, according to F's rules of inference, are compared to observables in N through a decoding represented by arrow d. When the predictions match the observations on N, we say that F is a successful model for N. Otherwise the entailment structures could not be brought into congruence, thus F failed to model N.

It is important to note that the encoding and decoding mappings are independent of the formal and natural systems, respectively. In other words, there is no way to arrive at them from within the formal system or natural system. That is, the act of modeling is really the act of relating two systems in a subjective way. That relation is at the level of observables; specifically, observables which are selected by the modeler as worthy of study or interest.

Given the modeling relation and the detailed structural correspondence between our percepts and the formal systems into which we encode them, it is possible to make a dichotomous classification of systems into those that are *simple* or *predicative* and those that are *complex* or *impredicative*. This classification can refer to formal inferential systems such as mathematics or logic, as well as to physical systems. As Rosen showed (1991, 2000), a simple system is one that is definable completely by algorithmic method: All the models of such a system are Turing-computable or simulable. When a single dynamical description is capable of successfully modeling a system, then the behaviors of that system will, by definition, always be correctly predicted. Hence, such a system will be *predicative* in the sense that there will exist no unexpected or unanticipated behavior.

A complex system is by exclusion not a member of the syntactic, algorithmic class of systems. Its main characteristics are as follows. A complex system possesses non-computable models; it has inherent impredicative loops in it. This means, it requires multiple partial dynamical descriptions—no one of which, or combination of which,

suffices to successfully describe the system. It is not a purely syntactic system as described by Eqs. 1–5 but it necessarily includes semantic elements. Complex systems also differ from simple ones in that complex systems cannot be linearly composed of parts—they are non-decomposable. This means, when a complex system is decomposed, its essential nature is broken by breaking its impredicative loops.

This has important effects. Decompositional analysis is inherently destructive to what makes the system complex—such a system is not decomposable without losing the essential nature of the complexity of the original system! In addition, by being not decomposable, complex systems no longer have analysis and synthesis as simple inverses of each other. Building a complex system is therefore not simply the inverse of any analytic process of decomposition into parts, i.e. the system is not a direct product of components, thus Eq. 5 does not hold.

Since the brain is a complex, integrated and thus non-decomposable system, both steps of reverse engineering—decomposition into functional and structural components and subsequent synthesis—must necessarily fail and will not provide the envisaged understanding!

It should be stressed that simple and complex systems after Rosen's definition cannot be directly related to those sensu Simon (Sect. 2). While Rosen's approach yields a *descriptive* definition of complexity, Simon's is *interactional*, see (Wimsatt 1972). It seems clear, however, that Rosen's 'simple systems' comprise Simon's full- and near-decomposable systems, and Rosen's 'complex systems' correspond to Simon's non-decomposable, integrated systems, as well as to Levin's evolved systems. No matter which definition is applied, the conclusion about the brain's non-decomposability remains valid.

5 Conclusions

If one attempts to understand a complex system like the brain it is of crucial importance if general operation principles can be formulated. Traditionally, approaches to reveal such principles follow the line of decompositional analysis as expressed in the *fundamental assumption* of cognitive and brain sciences, i.e. cognitive systems like other, truly complex systems are decomposable. Correspondingly, reverse engineering has been considered the appropriate methodology to understand the brain and to engineer artificial cognitive systems.

I have argued that this position is at odds with the findings of complexity science. In fact, non-decomposability *is* an intrinsic property of complex, integrated systems, and cannot be considered as subjective, temporary failure of our methodology, due to insufficient mathematical techniques. Thus, the dominant complexity concept of cognitive and brain sciences underlying reverse engineering needs revision. The updated, revised concept must comprise results from the nonlinear science of complexity and insights expressed, e.g., in Rosen's work on life and cognition. In the first line, this means that the *fundamental assumption* of cognitive and brain sciences must be abandoned.

Organisms and brains are complex, integrated systems which are non-decomposable. This insight implies that there is no 'natural' way to decompose the brain, neither structurally nor functionally. We must face the uncomfortable insight that in cognitive and brain sciences we have conceptually, theoretically, and empirically to deal with complex, integrated systems which is much more difficult than with simple, decomposable systems of quasi-independent modules! Thus, we cannot avoid (at least in the long run) subjecting research goals such as the creation of 'brain-like intelligence' and the like to analyses which apprehend the very nature of natural complex systems.

References

Arbib M, Érdi P, Szenthágothai J (1997) Neural organization: structure, function and dynamics. MIT Press, Cambridge

Atkinson AP (1998) Persons, systems and subsystems: the explanatory scope of cognitive psychology. Acta Analytica 20:43–60

Bechtel W, Richardson RC (1993) Discovering complexity: decomposition and localization as strategies in scientific research. Princeton University Press, Princeton

Bechtel W (2002) Decomposing the brain: a long term pursuit. Brain and Mind 3:229–242

Bressler SL, Tognoli E (2006) Operational principles of neurocognitive networks. Intern J Psychophysiol 60:139–148

Brodkin J (2009) IBM cat brain simulation dismissed as 'hoax' by rival scientist. Network World, Framingham

Brooks R (2001) The relationship between matter and life. Nature 409:409–410

Cummins R (1983) The nature of psychological explanation. MIT Press, Cambridge

Cummins R (2000) "How does it work" versus "What are the laws?": two conceptions of psychological explanation. In: Keil F, Wilson RA (eds) Explanation and cognition, MIT Press, Cambridge, pp 117–145

Dennett DC (1994) Cognitive science as reverse engineering: several meanings of 'top down' and 'bottom up'. In: Prawitz D, Skyrms B, Westersthl D (eds) Logic, methodology and philosophy of science IX, Elsevier Science, Amsterdam, pp 679–689

Dennett DC (1991) Consciousness explained. Little, Brown and Co, Boston

de Garis H, Shuo C, Goertzel B, Ruiting L (2010) A world survey of artificial brain projects PartI: large-scale brain simulations. Neurocomputing. doi:10.1016/j.neucom.2010.08.004

Edmonds B (2009) Understanding observed complex systems the hard complexity problem. CPM Report No.: 09-203

Forrest S (1990) Emergent computation: self-organizing, collective, and cooperative phenomena in natural and artificial computing networks. Physica D 42:1–11

Frégnac Y et al. (2006) Ups and downs in the genesis of cortical computation. In: Grillner S, Graybiel AM (eds) Microcircuits:

the interface between neurons and global brain function, Dahlem Workshop Report 93. MIT Press, Cambridge

Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc R Soc London B 205:581–598

Grillner S, Markram H, De Schutter E, Silberberg G, LeBeau FEN (2005) Microcircuits in action from CPGs to neocortex. Trends Neurosci 28:525–533

Gurney K (2009) Reverse engineering the vertebrate brain: methodological principles for a biologically grounded programme of cognitive modelling. Cognit Computat 1:29–41

Henson R (2005) What can functional neuroimaging tell the experimental psychologist?. Quart J Exper Psychol 58:193–233

Herculano-Housel S, Collins CE, Wang P, Kaas J (2008) The basic nonuniformity of the cerebral cortex. Proc Natl Acad Sci USA 105:12593–12598

Horton JC, Adams DL (2005) The cortical column: a structure without a function. Phil Trans R Soc B 360:386–362

Hubel DH, Wiesel TN (1963) Shape and arrangement of columns in cats striate cortex. J Physiol 165:559–568

Hubel DH, Wiesel TN (1977) Ferrier lecture: functional architecture of Macaque Monkey visual cortex. Proc R Soc Lond B 198:1–59

Levins R (1970) Complex systems. In: Waddington CH (eds) Towards a theoretical biology, University of Edinburgh Press, Edinburgh, pp 73–88

Le Novere N (2007) The long journey to a systems biology of neuronal function. BMC Syst Biol. 1–28

Maass W, Markram H (2006) Theory of the computational function of microcircuit dynamics. In: Grillner S, Graybiel AM (eds) The interface between neurons and global brain function, Dahlem Workshop Report 93. MIT Press, Cambridge, pp 371–390

Marom S, Meir R, Braun E, Gal A, Kermany E, Eytan D (2009) On the precarious path of reverse neuro-engineering. Front Comput Neurosci 3. doi:10.3389/neuro.10.005

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

Mountcastle VB (1997) The columnar organization of the neocortex. Brain 120:701–722

Price CJ, Friston KJ (2005) Functional ontologies for cognition: the systematic definition of structure and function. Cogn Neuropsychol 22:262–275

Rakic P (2008) Confusing cortical columns. Proc Natl Acad Sci USA 105:12099–12100

Rockel AJ, Hiorns RW, Powell TPS (1980) The basic uniformity in structure of the neocortex. Brain 103:221–244

Rosen R (1991) Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life. Columbia University Press, New York

Rosen R (2000) Essays on life itself. Columbia University Press, New York

Ross ED (2010) Cerebral localization of functions and the neurology of language: fact versus fiction or is it something else?. Neuroscientist 16:222–243

Schierwagen A (2007) Brain organization and computation. In: Mira J, Alvarez JR (eds) IWINAC 2007, Part I: Bio-inspired modeling of cognitive tasks. Lecture notes in computer science, Springer, Heidelberg 4527, pp 31–40

Schierwagen A (2009) Brain complexity: analysis, models and limits of understanding. In: Mira J et al. (eds) IWINAC 2009, Part I, Lecture notes in computer science, Springer, Heidelberg 5601, pp 195–204

Schierwagen A (1989) Real neurons and their circuitry: Implications for brain theory. iir–reporte, pp. 17–20. Akademie der Wissenschaften der DDR, Institut für Informatik und Rechentechnik), Eberswalde

Simon H (1969) The sciences of the artificial. MIT Press, Cambridge

Suykens JAK, Vandewalle JPL, Moor BL de (1996) Artificial neural networks for modelling and control of non-linear systems. Kluwer Academic Publishers, Dordrecht

Systems of neuromorphic adaptive plastic scalable electronics (SyNAPSE). DARPA/IBM (2008)

Szenthágothai J (1983) The modular architectonic principle of neural centers. Rev Physiol Bioche Pharmacol 98:11–61

Uttal WR (2001) The new phrenology. The limits of localizing cognitive processes in the brain. MIT Press, Cambridge

von Eckardt B, Poland JS (2004) Mechanism and explanation in cognitive neuroscience. Philos Sci 71:972–984

Wimsatt W (1986) Forms of aggregativity. In: Donagan A, Perovich AN, Wedin MV (eds) Human nature and natural knowledge, D. Reidel, Dordrecht, pp 259–291

Wimsatt WC (1972) Complexity and organization. Proc Biennial Meeting Philos Sci Ass 1972:67–86