General Principles for Brain Design

Brian D. Josephson

Cavendish Laboratory, J J Thomson Avenue, Cambridge CB3 0HE, UK.

Abstract. The task of understanding how the brain works has met with only limited success since important design concepts are not as yet incorporated in the analysis. Relevant concepts can be uncovered by studying the powerful methodologies that have evolved in the context of computer programming, raising the question of how the concepts involved there can be realised in neural hardware. Insights can be gained in regard to such issues through the study of the role played by models and representation. These insights lead on to an appreciation of the mechanisms underlying subtle capacities such as those concerned with the use of language. A precise, essentially mathematical account of such capacities is in prospect for the future.

Keywords: brain design, language, abstraction, representation, models, neurocomputational science.
PACS: 87.19.La**.**

1. PRINCIPLES AND THE BRAIN PROBLEM

This paper has as its aim the formulation of a set of concepts adequate to understanding the subtleties of the workings of the brain, including capacities such as natural language. Whereas in the case of the body in general the relationships between phenomenon and mechanism are typically of a transparent character, the same is so, in the case of the brain, only in comparatively simple situations. What we typically have there is either an account of a specific type of behaviour of a kind that is not transparently generalisable to cognitive functions in general, or alternatively a very general kind of theory (cf. Quartz and Sejnowski[1]) whose applicability to specific problems is equally unclear.

Is this lack of transparency fundamental to the brain, or is it simply that appropriate concepts are not being brought to bear on the problem? There are many cases in science where a single idea radically transforms the situation, allowing progress to be made in ways that previously were not possible. The way that this happens is that new concepts (e.g. the genetic code) are associated with generic models, instantiated in particular forms in particular systems. Once we understand the generic model, we are in a position to apply the concept in all situations where it applies, throwing light on situations that were previously incomprehensible. The expansion in understanding that results may lead to further insight based on additional concepts that come to light.

In the following, we gain access in this way to new ways of thinking about the brain, starting from the fact that while in principle the behaviour of an individual biosystem is derivable from fundamental quantum mechanics, this is not how we understand the behaviour of biosystems in practice (indeed, such a derivation for an individual biosystem will in general fail to inform us of the behaviour of the whole collection of systems of the type that is of interest, rendering such a first principles derivation of little value). Instead, we normally explain the behaviour of biosystems on the basis of descriptions at a higher level, involving for example specific molecules and specific chemical reactions, such derivations ignoring factors such as the precise positions of the molecules necessary for a first-principles derivation from quantum mechanics. Explanations of the workings of man-made mechanisms are similarly normally formulated in terms of high level descriptions involving entities such as levers or amplifiers.

It is reasonable to assume that a clear understanding of how the brain works, able to give a proper account of the subtleties of processes such as those of language, would similarly have as its basis higher level forms of description than ones at the neuronal level, since descriptions at that level fail to yield the desired understanding, the role of the neuronal level of description according to this point of view being largely limited to that of justifying the higher levels, in the same way that the laws governing the behaviour of amplifiers and levers are ultimately justified in terms of basic physical laws.

What is being asserted here is that there is a clear logic to processes such as language, not readily perceptible from information at the neuronal level. An analogy to this exists in computer programming, where it is similarly difficult to discern the logic that underlies the performance of the computer running the program by simply examining the microscopic details (in this case the executable code present in the computer while the program is running). It might indeed be considered irrational to imagine that the brain could behave in the way that it does without there being some kind of logical design of the kind hypothesised. The conventional way of thinking about the brain is the following: the observed behaviour follows from the nervous system architecture, while at the same time the latter is constrained by what is known, or given, concerning the behaviour types that are observed. In the alternative proposed here, the design plays the role of a logical intermediary: the design (i.e. the high-level account that governs the behaviour in an analysable manner) is the primary determiner of the behaviour, whilst also being an essential constraint upon the neural architecture.

However, merely hypothesising the existence of a higher-level of description of the mechanics of the brain does not in itself help us to a great extent with our explanatory problem. To be able to answer matters of detail (e.g. the organisation of the language system), rather than just giving explanations of a general kind, implies that the higher-level of description is also very complicated, leaving still uncertain the connection between the mechanics associated with the description and the behaviours of interest. The resolution of the dilemma consists in a consideration of design principles, which can be thought of as a deeper level of design underlying the kind of specific design specification considered so far. Design principles, unlike specific designs, have a generic character. As with concepts in general, they are associated with generic models, instantiated in particular forms in particular systems. Some generic models have properties that make them valuable in a biological context, leading to the widespread presence of systems conforming to these models.

It is our contention that such a state of affairs applies in the case of the brain, meaning that a number of principles, already known in different contexts such as computing, may be equally germane to the architecture of the nervous system. Nature has jumped ahead of our scientific understanding in making use of them. However, once we become aware of the principles and their applicability (see §6 for a general discussion of such a 'dual discipline', analogous to existing dual disciplines such as astrophysics and molecular biology) we can make use of this awareness to further our understanding of the brain, and will be able move beyond the relatively primitive level of understanding that we have achieved so far.

2. CONCEPTS FROM PROGRAMMING PRACTICE

Modern computing makes use of a number of technical devices or design principles, of which we shall focus here on the use in programs of a hierarchy of classes, each involving a collection of processes (known as methods or functions), with related parameters, working together in an integrated fashion[2].

Hierarchical design, involving the progressive definition of new functions in terms of ones that have already been defined in the program, has been utilised in programming from the beginning; with this methodology, the analysis of a complicated program breaks down into analysis of the individual modules involved in defining one function in terms of another, a considerable simplification that helps to ensure that a program will behave in the way required of it. The use of classes is based on the recognition that many types of situation are closely related to each other, allowing them to be treated identically (more precisely, on the basis of identical models), provided that any differences are taken into account by the use of parameters. This form of abstraction means that a single piece of code can do work that would normally require many, leading to increased simplicity and reliability. The practical implementation of the class concept involves the use of discrete units known as objects, with one object corresponding to each instance of the class, each object containing the data (possibly changing over time) that distinguishes the given instance of the class from the others. New objects are ‘created’ (in the form of blocks of memory containing the relevant information) when a new instance of a class is encountered.

We now consider the relevance of these techniques to the functioning of the brain. The idea of a hierarchy of processes is certainly relevant in the brain context, as is also the fact of the existence of generic classes of situation (e.g. balancing activity, getting from one situation to another, or using signs), handled in ways specific to the class; in the way discussed above, each instance of a class is distinguishable from the others by parameters specific to the class (e.g. different instances of getting from one situation to another being distinguished by parameters such as the start and end points, and important intermediate points on the route serving to define the route). The problem is how to relate these generalities in a practical way to the actual nervous system, including the desired transparency regarding the relationships between phenomenon and mechanism. This transparency, according to the present point of view, depends on the existence of a high-level description of what is happening, which itself achieves transparency by conforming to comprehensible design principles. Our hypothesis is that the similarity of this high-level account to accounts relating to computer programs means that its functioning would be accessible to means of

analysis familiar from computer programming, and desirable features possessed by computer programs, such as the simplifications associated with the use of hierarchies of classes as discussed above, would carry over to the nervous system case.

Arriving at this presumed high level account would require a detailed study of the systematics of behaviour, which would then have to be related to underlying neural mechanisms, a process analogous to that of biology generally, where one studies both biological processes and underlying mechanisms. In the present context, the processes involved relate to each other in ways analogous to the programming context. The way in which this is relevant to understanding processes such as language will be discussed later.

The above argument is a somewhat delicate one, amounting in essence to the possibility of utilising good ideas developed in one context in another, very different one. Modern programming depends on certain ideas which find expression in the source code of an appropriate programming language, which language incorporates these ideas directly. The question then has to be addressed: how does the nervous system enact these programming concepts, upon which enaction, according the present point of view, the cognitive abilities of the nervous system critically depend? The answer, in general terms, is the following. In the computing case, enaction of the high-level source code is implemented automatically by the compiler for the language concerned, whose design is essentially a matter of problem-solving, once the meaning to be attached to the terms of the language has been precisely specified. A similar problem-solving exercise would be needed to enact the high-level prescriptions in the nervous system context, the requirement that nervous system architectures must exist to realise the prescriptions of the high-level account constraining what these prescriptions may be allowed to assert that the system should do.

In other words, the nervous system needs to have correlates to, or substitutes for, the basic processes available in the context of programming languages. Neural mechanisms are related to mechanisms used in digital computation in the same sort of way that mechanical switchboards and electronic switchboards are related. The brain has to make use of an ancient technology to carry out specific operations that would be far simpler to implement using digital computers. Typical requirements are those of representation, information storage, and memory allocation.

Representation is perhaps the most fundamental of these, since nothing can be decided until it is determined what should correspond physically to the elements of the high-level description. Subsequent to this, the basic requirement must be that the physical processes correspond, under the terms of this representational scheme, to the processes prescribed in abstract terms by the high-level description. Neuroscience tells us that the brain does indeed utilise specific representational schemes of a generic kind, with processing by neurons carrying out the required operations. In general, such representational schemes need to include what is characterised in the programming context as data format. Variables in computing often have a complicated structure, represented in memory in accord with prescribed rules, and our picture requires, in the same way, that structured types of information be represented in the brain in accord with specific rules, which rules have to be taken into account in the realisation in the architecture of the dynamical processes in which the structured information participates.

The account given so far fails to address the precise aspects of functionality that are observed, which precision cannot be built in, in the way that it can be built in with artificial mechanisms. But if the relevant models for the process include learning on the basis of systematic trial and error, they can account for the development of accurate functionality (learning has an unusual role in the nervous system context since neural networks have been found to be better than conventional algorithms at certain learning tasks. We deal with this situation by allowing the high-level specification to include learning operations, the corresponding analysis being based on an appropriate model of the results of the learning process).

We consider now in more detail the role of models in accounting for functional behaviour. In the computing context, as already observed, models are used to verify that the code for a given section of program correctly generates the prescribed behaviour for the given functions and, as discussed in connection with the notion of classes, to a first approximation instances of a given class can be treated identically provided that any differences are taken into account by the use of parameters, so that a single piece of code can accommodate a range of situations, leading to increased simplicity and reliability. Here, analogously, a single prescription, implemented by a single circuit for the given class, can do work that would otherwise require many. The nervous system needs to have a neural equivalent to the process of allocating a new block of memory for a new object in computer programming, and information acquired during learning activity associated with an instance of the class (driven by the single circuit for the class as discussed) should modify the module associated with the class instance, invoking the learning mechanisms in that module in a way that accords with the associated formatting conventions. This circuitry is likely to include mechanisms designed to stay in a given situation for some period while the relevant learning takes place.

Throughout the above, it is assumed, as discussed, that special mechanisms exist in the brain to carry out subtasks needed in computer programming, providing for example equivalents to putting information into memory, allocating memory for modules, etc. For example, interconnections between neural units can play the role of

pointers (i.e. the storage of a specification of one memory location in another memory unit) in conventional programming.

This is as far as it is possible to take these arguments at this point, except to make the general point that an essential corollary of the above line of argument is that features of the neural circuitry that seem to have no clear significance according to current thinking are likely to be in fact implementations of the very specific aspects of the complicated overall design that is suggested by this new approach. Here all that can be done is to point out general principles. Understanding fully how they are applied in the actual brain will be a task of similar magnitude to the development of modern computing from the pioneering efforts of Babbage and Turing.

3. RELIABILITY: THE ESSENTIAL QUEST

The brain has to contend with somewhat different circumstances than most computer programs. While man-made mechanisms are, as far as possible, designed so that once any set up processes have been carried out the mechanism is functional right away, this is not the case for biosystems, where in most cases a learning period is necessary to achieve effective functionality. The difference is a corollary of the fact that man-made devices can generally perform in accord with a precise specification, while in addition the environment of artificial devices often conforms accurately to specific laws, as a result of which an appropriate model can be used to create a fully-functioning design. In the biological case, such precision is not possible, but nevertheless reliable functioning may be achievable by means of systematic trial and error, and appropriate adjustment. Thus qualitatively different models must apply to the design on which the nervous system is based, ones concerned with trial and error in ways that a certain fraction of the time achieve reliability in particular things. Since such reliability cannot be assured, some mechanism, probably based on inhibition, is required to ensure that it is only the more reliable systems that are, wherever possible, the ones used in subsequent activity. The discovery and development of reliable systems of various kinds, through a succession of constructions in the style of the Baas hyperstructure theory[3], is crucial to the way the system works as an integrated whole, the models on which the design is based needing reliable components in order that they be applicable. Note that this integration has a top-down aspect through the way that, through the models, the behaviour of wholes can be fed back to modulate 'intelligently' the behaviour of the component parts.

We can now see in principle how processes such as those associated with language can work so effectively. Suppose that the usage of language involves a range of generic devices characteristic of language, each handling an aspect of language and each governed by an appropriate model. With each model is associated a mechanism for achieving reliability in regard to the function concerned (e.g. determining the meaning to be ascribed to a word, or the way in which a particular grammatical construction is used in a given context). Whether a given application of language is successful is a function of the reliability of the individual components. In alternative terminology, reliability relates to linguistic competence, which manifests itself in the form of successful linguistic and post-linguistic performances. From this perspective, a language system is a complex system developed over time by 'intelligent' trial and error, retaining what works reliably. A precise analysis of the models relevant to the various levels, characterising what amounts to intelligent choices in the various contexts, should make possible detailed explanation of the various functions of language. More qualitatively, the ideas developed here can be exploited to understand the mechanisms whereby our cognitive domain achieve success in more and more abstract realms as time goes on, through the initial discovery, followed by the extension of, such possibilities.

4. COGNITIVE CATEGORIES IN DESIGN

Special cognitive categories, such as those concerned with language in the above discussion, or with aspects of spatial cognition, emerge in a natural way from our approach, through the connections between models and design, since specific kinds of information enter into the individual models. For example, the concept of position relative to the body enters in models as something that can be inferred from visual information, as well as entering into specifying an action (e.g. grasping an object at a particular location). Thus we envisage that the appropriate neural systems begin to represent, in this example, position, when, in the course of development, they attempt to execute a task where the relevant model involves that particular entity. In the development, some neural subsystem in the appropriate area would be selected by the hardware configuration to represent the position of current relevance, and connected with other systems in ways consistent with the model governing the relevant task. Thus that particular position (e.g. relative to the body) would become a 'familiar position', and over time a comprehensive stock of such familiar positions would be accumulated, in the same way that one builds up a stock of familiar people or familiar words, each with its own attributes characteristic of its class.

In the present approach, the details of design enter naturally, rather than in an ad hoc manner, though the existence of models relevant to effective functioning, entailing an account that is in essence mathematically grounded. Naturally, experimental information will be needed, in addition to theory, in order to determine which of various possible designs corresponds to the one that actually chosen in the course of evolution.

5. THE THEMATICS OF ABSTRACTION AND LANGUAGE

While a detailed analysis lies far beyond the scope of the present work, some interesting possibilities tentatively suggest themselves. We can, for example establish links between the present approach and ideas developed by workers such as Karmiloff-Smith[4], Jackendoff[5] and Arbib[6], offering the possibility of putting such proposals on to a firmer footing. Karmiloff-Smith, for example, proposes that there are specific levels of representation associated with different levels of abstraction. The working out of such a proposal, according to the present scheme, would involve very specific information-acquiring cognitive actions (e.g. to determine whether a given object should be considered an element of a specified class), grounded in very specific models concerning the relationships between the contents of the different levels. Again, language, the logic of which has been analysed in detail by Jackendoff and also by Arbib, can be seen as having its own characteristic models reflecting possible uses of language and attributes of language use. Our approach would systematise this work by hypothesising that the basis of all developments in language is the investigation of the possible fits between aspects of a current situation and a generic model, leading in ways specified by the model to the creation of a reliable component of the language system. These would lead to the building up of a range of higher level processes, in ways that could be modelled in detail.

Any given language system depends on what has been discovered about the possibilities of language by language users, discoveries that are constrained by general principles such as those studied in linguistic theory. For example, the fact that in some circumstances the future is partly predictable, allied to the possibility of characterising such predictions using language, may lead to the discovery by an individual that language can be used for such a purpose, which knowledge may then propagate throughout a language community. There are very many other possibilities of this kind.

There are two issues in particular where the viewpoint developed here may be able to account for important features of language. An example is the existence of categories such as NP and VP (noun phrase and verb phrase). These are puzzling since they do not correlate completely with the cognate semantic categories of object and action. The explanation may reside in the fact that the primitive distinction between whether a language user is attempting to indicate an object or an action may have led to the evolution of a corresponding differentiation, in the neural hardware, in the neural representation of signs. More advanced levels of linguistic activity may then have developed to take advantage of this differentiation, e.g. to detect phrase structure, which differentiation would lose its semantic import since the semantic distinctions involved could at that point be indicated in alternative ways.

Another issue is the presence and utilisation of syntactic structures that are perceivable as tree-like, i.e. the phrase structures just alluded to, the exploitation of which, as discussed by Jackendoff and by Arbib, is associated with a major advance in the evolution of language. From the present point of view, tree-like structures could emerge naturally through the process of organising the structure of speech, as a consequence of this being a characteristic of action in general. The listener's job is to determine 'where the structure is coming from', which may be facilitated by the existence of a 'mirror system', along the lines proposed by Arbib, to represent actions in such a form as to facilitate imitation.

This implies an interesting kind of situation, where a combination of linguistic activity (encoding information in speech) and a non-linguistic mechanism (the generation of a structure in principle perceivable as a tree) generates a task of specifically linguistic import (discovery of the structure that led to the generation of the tree, whose form bears a general relationship to that of the tree). This would motivate the evolution of a system specialised to this specific task. Such a system may, as Arbib suggests, be related to the activity of planning a grasping action, but it is likely that some aspects of the details are language-specific, or at least connected with the skill of 'mind-reading', itself in principle an offspring of the capacity to imitate governed by its own characteristic models. The question remains for the future what the specifics of the present approach may have to say regarding the very specific features of language used by Jackendoff to argue for very specific features of the language system.

This concludes our analysis, where a very general vision of what underlies the effectiveness of the brain has led us on a complicated trail. Many technical details are involved, but this should be unproblematic since it is the essence of biosystems to discover and exploit technical devices. It is to be hoped that this vision reflects the reality sufficiently well that it can be of value in clarifying the puzzle of the brain for the future.

6. A NEW DISCIPLINE: NEUROCOMPUTATIONAL SCIENCE?

The above discussion can be thought of as the indication of steps towards the initiation of a new discipline, which might be called Neurocomputational Science. As with existing dual disciplines such as astrophysics and molecular biology, this process would involve bringing together two distinct areas of study, in this case neuroscience and computer science. Commonalities between the neurosciences and conventional computing have long been recognised, but only in regard to simpler aspects such as information representation and processing, and algorithms. Our position is that purposive computation has its own logic, with certain consequences that are independent of the underlying hardware. By identifying the points of contact between computer science and neuroscience, and understanding the relationships between the two types of situation, should greatly enhance our understanding of the latter, in the same way that physics enhances astronomy and chemistry enhances biology. Future publications will develop this point of view in detail.

ACKNOWLEDGMENTS

I am indebted to Profs. Nils Baas and Andrée Ehresmann for stimulating discussions.

REFERENCES

1. S. R. Quartz and T. J. Sejnowski, *The neural basis of cognitive development: A constructivist manifesto*, Behavioral and Brain Sciences, **20(4)** 537-596 (1997).
2. *The JAVA Tutorial*, http://java.sun.com/docs/books/tutorial/java/concepts/ (2005)
3. N. A. Baas; Emergence, "Hierarchies and Hyperstructures," in *Artificial Life III* edited by C.G. Langton, Redwood City, CA: Addison-Wesley, 515-537 (1994).
4. A. Karmiloff-Smith, *Beyond Modularity: a Developmental Perspective on Cognitive Science,* Cambridge, MA: MIT Press (1992).
5. R. Jackendoff, *Foundations of Language*, Oxford: Oxford (2002).
6. M. Arbib, "The Mirror System, Imitation, and the Evolution of Language," in *Imitation in Animals and Artifacts*, edited by C. Nehaniv and K. Dautenhahn (2000), Cambridge, MA: MIT Press.