Abstractions and the Brain

Brian D. Josephson
Department of Physics, University of Cambridge
Cavendish Lab. Madingley Road
Cambridge, UK. CB3 OHE
bdj10@cam.ac.uk
http://www.tcm.phy.cam.ac.uk/~bdj10

ABSTRACT

The following is the handout that accompanied a paper presented at the December 2001 Messina conference on Horizons in Complex Systems. Currently it presents the basic concepts only. It proposes that one key idea constitutes the key to understanding the brain, namely the fact that abstractions are possible. The particular abstractions relevant to particular aspects of nature define the designs that are capable of handling these aspects of nature. The task of explaining the brain thus reduces to an investigation of the abstractions, the relevant interrelationships, and the corresponding design components.

1 BRAIN FUNCTIONING

This paper is concerned with understanding how the highly complex structure that is the human brain is able to accomplish advanced skills such as using language. Of the existing approaches to understanding nervous system functioning, one important one is that of experimental studies of brain and behaviour, and another that based upon computational models of neural networks (e.g. Elman et al 1996). Neither offers insights into the subtleties of skills such as language, the former because neural circuitry can account for such behaviour in qualitative terms only, and the latter because the behaviour that it has been practical to simulate involves only rather basic aspects of linguistic behaviour and it is not at all clear how significantly more complex aspects of language are to be modeled. A third approach is that of Minsky's society of mind (1987) which involves discussion of ways in which neural networks could emulate the kinds of behaviour exhibited by conventional computer programs. Such programs are able to model complexities of behaviour to a certain extent, but the approach suffers from two drawbacks, firstly in that conventional computer programs do not provide a good model for brain processes generally, and more seriously in that the developmental processes which lead to the acquisition of skills are discussed in the model only to a very limited degree.

1.1 ABSTRACTION

The following approach, inspired mainly by the ideas of workers such as Baas (1994) on hyperstructures and the observer mechanism; Ehresmann and Vanbremeersch (1987) on relational aspects and Karmiloff-Smith (1992) on experimentally motivated concepts such as domain-relevant activity and representational redescription, has a very different character. It focuses on the relevance of abstractions and relationships to matters of design. A simple illustration is provided by Ohm's law, V = IR, where V, I and R denote the voltage across a resistor, the current through it and the resistance respectively. Here the entities symbolised by V, I, and R are abstract entities in a scheme embodying one relationship, namely that specified by Ohm's law. A physical resistor that satisfies Ohm's law provides a realisation of the abstract scheme. Designs make extensive use of realisations of abstract schemes such as resistors and microprocessors since they can utilise the properties that such systems possess by virtue of being such realisations. Properties associated with realisations of particular abstract schemes conversely feature in explanations of designs; in practice, labels or descriptive accounts are used to indicate that particular schemes apply.

In cases such as that of a resistor, the fact that a given abstractional scheme applies is known through experiment or physical theory, but in the microprocessor case it is known through logical inference, the properties of the components in accord with the schemes that are assumed to apply to them implying the properties of the whole. Thus in a reductionist analysis, 'conformance to an abstractional scheme' is something that propagates upwards and allows us to infer in appropriate cases how highly complex systems should behave. Ensuring that such inferences are valid is the essence of design, which consists of a list of abstractional schemes, combined with a specification of the mechanisms that ensure conformance to them.

The same ideas apply equally well, but in a less rigorous sense and as an idealisation, to biosystems. Like mechanisms, they contain components of various kinds, each type conforming to some scheme of abstractions that corresponds to our understanding of the types of entities concerned. The design aspect consists of the various mechanisms that help the systems concerned conform to their particular abstractional schemes. Biosystems differ from machines in that the entities concerned often lack a formal specification, their properties being inferred from investigations of instances of the entities concerned that are encountered in nature. The inferences involved in going from one level of description to another are similarly typically nonrigorous, being based instead on a range of ideas justified in various ways. What makes this a scientific process rather than mere guesswork is that the various assumptions made are

open to experimental testing and, where appropriate, refinement and replacement by a better account.

It is reasonable as a working hypothesis to postulate that explanations of the same general character would apply equally well to nervous system functioning. The implication is that the nervous system, in its environment, is capable of being characterised as a hierarchy of systems conforming to a range of abstractional schemes, the design being such as to cause the systems concerned to tend to conform to the various schemes. This characterisation can be usefully compared with conventional computing systems, which also depend on systems that conform to specified abstract schemes, such as one whereby sending a code for a character to the relevant system leads to the character concerned being displayed on the screen. The difference between the brain and the computer is that in the case of the computer the systems concerned are defined directly by the (compiled) program, whereas in the case of the brain most of the systems conforming to given abstractions are created through the process of development, the design thus determining the details of the system indirectly, rather than directly as in the case of a computer program.

The abstractions we are concerned with typically relate to particular neural circuits or systems and their behaviour in a given environment, and are thus similar to abstractions relating to computer software. The existence of such systems, logically interrelated in various ways leading to explanations of complex behaviour, is our key assumption. Their existence is taken to be the product of an effective design, consequent upon the processes of evolution, embodying a range of generative systems that themselves bring such derivative systems into existence in the course of development or learning. Examples are generative systems for acquiring the ability to maintain balance, for taking steps, or for defining routes.

This assumption is similar to Karmiloff-Smith's (1992) concept of modularisation, differences lying in the additional fact that the detailed design of the hardware concerned is taken here to be governed by abstractional schemes, and also the idea that modularisation can be effective at a number of levels. The logic of the link between design and abstractional schemes is that effective designs are grounded upon theory, while theories are formulated within abstractional schemes. The multilevel capabilities associated with abstractional schemes, on the other hand, involve in essence the fact that one mathematical system can contain entities on which another system can be based, just as when for example we extract out of the set of all transformations the subset consisting of all linear transformations, a collection that is associated with mathematical schemes of its own. The application to cognitive processes is that a developmental process may have its eventual outcome 'target processes' subject to their

own simplifying abstractions. For example, one aspect of learning to walk consists in learning how to walk directly to a visible destination. This outcome has a particularly simple abstract specification that can form the basis of higher capacities such as going to a more distant location indirectly via a series of intermediate destinations. The abstractional scheme concerned with the latter is concerned issues as the direct accessibility of one point on a sequentially defined route from the previous one.

One can go into the question of design for a specified result more deeply, while still talking in general terms, by noting (a) that the links and neural processes in a neural circuit define relationships while (b) that all relationships associated with a circuit are determined by the basic relationships of

(a) Changing one of the basic relationships has a specified effect on all other relationships, in principle allowing the existence of mechanisms for creating a system conforming to some target condition in a systematic way. The successful designs are ones that achieve this.

The above is not intended as a statement as to what a successful design is, rather it is a clarification of how successful designs work, an essential to the understanding of how the concepts developed here may be utilised to make sense of the complexities of the brain, the key to the latter being to use the information available to determine what are the abstractions on which are based the various components of the design.

1.2 LANGUAGE

Finally, we return to the issue with which we began, that of the processes associated with language, where it is controversial whether there are specific mechanisms for language (the nativist claim, connected with the existence of linguistic universals), or whether language abilities come about as a result of general learning mechanisms in an environment where language is present (the constructivist hypothesis), or some intermediate hypothesis. The present picture leads us to hypothesize that the design of the brain is linked to a number of abstractions related to language, use of which facilitates development of the capacity to use language. There is a connection with the work of Pinker (1994) who discusses regularities of language related to its effectiveness, and proposes that innate mechanisms mediate these regularities. We also make use of Karmiloff-Smith's (1992) concept of representational redescription (RR), and begin our account within that framework, according to which information is represented in a number of different formats at different times, a more advanced format coming into play subsequent to a more elementary one having been mastered in the given context. This idea can be usefully related to the abstraction of equivalence, whereby different means may be available for representing the same information,

which differ from each other in regard to particular characteristics and in the ways in which they may be used.

In Beyond Modularity (1992), Karmiloff-Smith discusses in considerable detail how the RR scheme can be related to observations of development. In the following we focus instead in very general terms how it can be related to the functioning of language. An important concept is the following: from an existing representation A, valid in situation S, there may be developed a different but provisionally equivalent alternative mode of representation B. The data a and b in representations A and B are related within some abstractional scheme, which defines the design of the system that generates b from a. This system may include a part that verifies the equivalence of A and B according to the scheme. One may then try to find something in a new situation S', a', say, which is operationally equivalent to b in the new situation (and so indirectly equivalent to a). Thus with appropriate criteria for equivalence it may be possible to adapt the action in situation S to a new situation. The same representation b applies to both activities so it may be regarded as a generalisation.

Thus, activity is developed on to a more abstract plane. It may be extended over time to the activity of planning, where one develops processes at the B level that are equivalent to those at the A level. Equivalence can then be used to try out a process at the B level before enacting it at the A level.

Such processes can now be envisaged at a more subtle level, C say, where the representations are of a more symbolic character, including in particular symbols for relationships. In other words, relationships which were explicit at say the B level are indicated in accord with an associated token at the C level. The explicit-symbolic relationship is itself an abstraction that can determine the design of circuitry to implement it. Such more abstract representations can be investigated for their utility and used to expand the possibilities further.

Language is a more subtle level again, characterised by the fact that it involves coding processes, or equivalently procedures for defining equivalence, that can be adapted to needs. The system derives its power from the fact that it embodies a range of options for linking strings of signs to various powerful representations at other levels. The development of a language is in essence the trying out of various possibilities with the exploration of what they can do. One possibility is simply the assignment of a name to something, and another the linkage of particular forms at the language level to forms at other levels according to a specific rule, these two being the main basis of the expressive power of language according to Pinker. These processes can be accommodated within particular abstractional schemes related to universal

grammar, which determines what kind of neural circuitry could implement such schemes.

In more detail, language is assumed to be based upon the equivalence of information represented as language and information expressed in other levels. Equivalence is a matter both of definition (and the operations of the brain's translation mechanisms for determining equivalence) and of the pragmatics of language as a communication. In other words, language use presupposes that a listener will generate an equivalent and be predisposed to act as if the information came from a different source, this providing a test for whether the translation was done correctly. In other words, correct translation should generate an 'idea' that fits the demands of the current situation.

The question now is whether such ideas are sufficient to generate something like language as it occurs naturally. This requires in particular correct syntactic analysis and the creation of the appropriate corresponding data structures. The answer that one would hope for would be the case is along the following lines. A language system (or more accurately the users' linguistic processes) defines certain equivalences that form the basis of its use. Comparatively simple cases allow users to determine which equivalences are part of the language and build up their own translation systems (on the basis of mechanisms adapted to the various kinds of abstractions involved in the equivalence).

Through the use of devices such as working memory, these systems can handle complex language equally well, but increased complexity brings more risk of error. But language users adapt their use of the relevant systems so as to minimize the risk of error, thereby continually increasing the possibilities of the language system. These considerations apply equally to pragmatic use of language (the use of language to achieve particular goals) and to the complexities of the language system itself.

A technical aspect of language is the conversion from linear strings to hierarchical structures which, as is well known, is connected with the ability to detect a valid group and 'iconise' it as a single entity, forming a node of a tree. This detection is based on pattern detection, itself utilising categories, some of which appear to be innate. Innate categories are in principle expected on in the present picture, assuming that they feature in some of the abstractional schemes, thus being expected to have correlates in the neural hardware.

2 SUMMARY

This completes our discussion, which is of a tentative character. A principle has been established involving general connections between abstractions and design. Since

abstractions of many kinds appear to feature in how we perceive and understand the world, and the organisation of the nervous system appears to reflect such abstractions, it is tempting to see this as a fundamental principle behind the workings of the brain, exploitation of which will radically advance our detailed understanding of how it works.

ACKNOWLEDGEMENTS

I am grateful to Professors Nils A. Baas and Andrée Ehresmann for numerous discussions, which assisted in the formulation of the above ideas.

REFERENCES

Baas, N.A. 1994. Emergence, Hierarchies and Hyperstructures. In: *Artificial Life III* C.G. Langton. (Ed.) Addison-Wesley. Pp. 515-537.

Ehresmann, A.C. and Vanbremeersch, J.-P. 1987. Hierarchical Evolutive Systems: a Mathematical Model for Complex Systems. *Bulletin of Mathematical Biology; Vol. 49, No. 1*. Pp. 13-50.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A.; Paresi, D. and Plunkett, K. 1996. *Rethinking Innateness: A Connectionist Perspective on Development.* Cambridge, MA: MIT Press.

Karmiloff-Smith, A. 1992. *Beyond Modularity: a Developmental Perspective on Cognitive Science*. Cambridge, MA.:MIT Press.

Minsky, M. 1987. *The Society of Mind*. Heinemann.

Pinker, S. 1994. *The Language Instinct: the New Science of Language*. Penguin.