September 1994 LIDS-P-2265

Autonomy and adaptiveness: a perspective on integrative neural architectures

Bernard Gaveau* **Charles Rockland**** **Sanjoy K. Mitter*****

* Universit é Pierre et Marie Curie (Paris VI), Mathématiques, tour 45-46 (5eme étage), 4 place Jussieu. 75252 Paris Cedex 05 FRANCE

** Center for Intelligent Control Systems and, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 35-304, Cambridge, MA 02139-4307 USA

*** Department of Electrical Engineering and Computer Science and Center for Intelligent Control Systems. Massachusetts Institute of Technology, 77 Massachusetts Avenue, 35-308. Cambridge, MA 02139-4307

This research has been supported by ARO Grant DAAL03-92-G-0115 (through the Center for Intelligent Control Systems), by the Scuola Normale Superiore, Pisa, Italy through funds provided by IBM-Semea, Italy and by a contract under EU "capital humain et mobilité".

Acknowledgements:

Bernard Gaveau wants to thank the Laboratory for Information and Decision Systems at MIT and the Scuola Normale Superiore for their hospitality.

Charles Rockland thanks the Université Pierre et Marie Curie and the Scuola Normale Superiore for their hosptiality.

TABLE OF CONTENTS

1. Introduction 1

1.1 Adaptive autonomous systems
1.2 The role of mathematical approaches
1.3 Organization of the paper

2. Various examples of complex systems and traditional approaches 6

A. Complex interactive systems 6

2.1 Computer-like systems
2.2 Various networks, manufacturing control process,...
2.3 Physical systems
2.4 Biological systems, decision systems
2.5 Other distinctions: natural vs. artificial, embedded vs. simulated
2.6 A range of metaphors

B. Some approaches to complex interactive systems: engineered systems 12

2.7 Approaches to be considered
2.8 Internalist vs. externalist perspectives
2.9 Symbols as emergent vs. imposed
2.10Emergence of behaviors
2.11 Combinatorial explosion of the state space
2.12 Control and coordination of behaviors
2.13 Motivation for action
2.14 Coherence
2.15 Non-algorithmicity

C. Neurobiological approaches to integrative organization 21

2.16 Approaches to be discussed
2.17 The theory of neuronal group selection
2.18 Convergence zones and somatic markers
2.19 Neural symbolization

3. Characteristics of Adaptive autonomous systems: what must a mathematical framework accommodate? 26

A. Characteristics of adaptive autonomous systems 26

3.1 The basic requirement: relative adaptation to the external environment
3.2 Basic constraints
3.3 Goals, values, and optimality
3.4 The need for control: achieving and maintaining coherence
3.5 Whate else is needed?

B. What can one expect from a mathematical theory of adaptive autonomous systems? 35

3.6 Is such a theory possible?
3.7 Historical background
3.9 Whate kind of methematical approach do we envision

4. Descriptions and prescriptions for adaptive systems 45
A. Constraints and specifications 45
4.1 Necessity vs. choice
4.2 Constraints
4.3 Specifications for an artificial system
4.4 Specifications for a natural system
4.5 Examples of specifications and constraints
B. Embeddedness and failure 56
4.6 Stances towards system failure
4.7 The problem of failure for simulated systems
4.8 Inadequacy of algorithmic models
4.9 The problem of failure for artificial systems
4.10 The problem of failure for natural systems
4.11 Concluding remarks
5. Some intentional notions for adaptive autonomous systems: behaviors, goals, and values 69
A. Actions and behaviors 69
5.1 Actions and elementary actions
5.2 The intentional notion of a behavior
5.3 Behaviors vs. classes of sequences of actions
5.4 Behaviors as emergent
B. Goals and values 75
5.5 The intentional concepts of goals and values
5.6 Primary goals and values
5.7 A basic "strategy" of adaptive autonomous systems: symbolization
5.8 Higher order goals or values and their symolization
6. Some basic analytical notions for autonomous system description 81
A. States and modules 81
6.1 States for Turing machines and for abstract models of computation
6.2 Modules
6.3 Finding relevant notions of "state"
6.4 "States" linked to the value system
B. Control: Maintaining Coherence 89
6.5 The need for control
6.6 "Principles" for control design
6.7 Current approaches to design of control laws: state-based and behavior-based models
6.8 Control of communicating modules
6.9 Control of interface relations via symbolization
6.10 Learning as control and learning to be controlled
7. Neural Symbolization 103
A. Characteristics of neural symbols 103
7.1 Physical description and role of neural symbols
7.2 Comparison and contrast with the notion of "symbol"

in AI and linguistics
7.3 "Homogeneity" of neural symbols
7.4 The content of a neural symbol
B. Encoding 110
7.5 The combinatorial strategy
7.6 The necessity of encoding by symbolization
7.7 Example: a schema for associations
C. Maintaining autonomy 114
7.8 Preservation of autonomy and relative adaption with the environment
7.9 The general symbolization strategy
7.10 Symbolization by a subsystem
7.11 Decision-making procedures
7.12 Avoiding a regress to infinity in decision-making
D. Categorization 125
7.13 Preliminary remarks: the process of categorization
7.14 A schema for categorization: modules observing other modules
Figures 132

1. Introduction

1.1 Adaptive autonomous systems

The present paper is an exploratory study directed towards a long-term aim, the development of a mathematical framework for the study of adaptive autonomous systems, with an emphasis on the integrative organization of these systems. We wish, as part of this program, to develop a language permitting us to attach meaning to the notion of "rationale" for the system's organization. The framework should provide, in the setting of integrative organization of autonomous systems, notions that can assume the role played by "laws of nature" in the physical sciences.

We address these questions both in the settings of natural and of engineered systems. We are not claiming that there is any fundamental congruence between these two[1] classes of system. However, we do feel that clarification of the contrasts between their distinct modes of organization may well facilitate the analysis of each. In addition, we expect that the design of engineered systems which are both adaptive and autonomous (to the extent that this is possible), will require significant insight into the principles of organization of natural systems, as well as significant reinterpretation of what is meant by the "design" process.[2] We have in mind both "small and "large" engineering systems. Some examples are: the design of an autopilot for a modern aircraft (a "small" system); defining the design, and understanding and structuring the design process, of a new aircraft (a "large" system); the design of a global communication network.

A word as to our use of "integrative". Here we do not insist on a "total", top-to-bottom integrative treatment incorporating all the organizational levels in the system (nor every level of detail) concurrently. Rather, the emphasis is on issues associated with maintaining coherence among multiple functionalities or behaviors. In particular, an "integrative" treatment in this sense in no way prejudices us against contexts where only schematic vs. detailed treatments are currently feasible (a case in point being higher cognitive function).

The paradigmatic autonomous systems are biological organisms, including single-cell organisms such as *E. coli* . Typically, one has in mind here organisms endowed with nervous systems, ranging from "simple" systems such as nematodes or insects, to human beings, with all their higher cognitive faculties, including language. (In fact, the starting point of the present study was a cognate research program, initiated by of one of us, taking as a model the nematode *C. elegans* .[3] This organism displays a rudimentary

---

1 Nor are we suggesting any monolithic uniformity within the "two" classes.

2 When we speak of "principles of organization" we mean this literally. In particular, we mean to distinguish this from simply the transposition to an engineering setting of "mechanisms" or "design tricks" drawn from a biological control system.

3 Rockland, C. (1989) The nematode as a model complex system: a program of research. Laboratory for Information and Decision Systems, Working Paper (*LIDS-WP-1865*) 1-115 (plus appendices).

Rockland, C. and S.G. Rowley (1990) Simulation and analysis of segmental oscillator models for nematode locomotion, MIT Center for Intelligent Control Systems Publication (*CICS-P-212*).

"intelligence", despite the fact that its "brain" contains only about three hundred neurons, and its whole body only about one thousand cells.)

In what sense are these organisms "autonomous" ? The construction of a formal characterization of "autonomous systems" is part of our long-term goal. Informally, however, one can point to several salient attributes of such systems: They must maintain themselves in their environment without the benefit or interference of external intervention. They must, accordingly, set their own "goals", select the "problems" to be solved, determine which external stimuli are relevant, and "decide" on choices of action. They may need to adapt to changes in their environment (and, to some extent, choose their environment), though they may (and in fact they ultimately do) "fail". All this may involve learning in various forms.

If we loosen, or relativize, our notion of "autonomous" (or consider subsystem organization), the systems of interest include individual cells or organized functional systems of cells. This includes, in particular, the various (quasi-independent or interdependent) control systems of very complex organisms, such as the nervous system, the immune system, etc.

Analogous questions can be addressed in settings not involving individual autonomous systems, but societies of quasi-autonomous systems (not necessarily "rational agents") , the society being itself regarded as an adaptive autonomous system, though at a different "hierarchical" level. Three examples are:(i) a large organization, such as a manufacturing firm; (ii) an economy (national or multi-national); (iii) an ecology. We think that various of the ideas raised in the present paper may be relevant to such systems.

In seeking to accommodate such systems within our range of interest, we are not being overly simplistic, nor engaging in a quixotic search for a "theory of everything". We recognize that not only at a detailed, but even at a highly schematic level, each such system is highly distinctive. It is not a case of "one theory fits all". However, at a minimum, the notion of "society" in the above sense, may be a highly useful metaphor (e.g., to be used as a counterpoint to the "machine" or "computer" metaphors) even in the study of a single biological organism. Two cases in point: (i) A cell may, from one natural perspective be regarded as a manufacturing plant. (ii) The mind of a single individual may be viewed, in various ways, as a society.[4]

Rowley, S.G. and C. Rockland (1991) The design of simulation languages for systems with multiple modularities, *Simulation* ,.56:3, 153-163.

[4] Two quite distinct examples in the psychology literature are: M. Minsky, "The society of mind", Simon & Schuster, 1985; G. Ainslie, "Picoeconomics: the strategic interaction of successive motivational states within the person", Cambridge U. Press, 1992.

Minsky presents a schema in which individual intelligence arises from the interaction of a society of "agents" which may themselves be without intelligence in any sense.

Ainslie's theory stems from a study of why individuals (both lower animals and humans) "irrationally" act upon temporary preferences for objectively inferior (and even harmful) goals. Anslie concludes (see p. 362): " The fundamental insight of picoeconomics is that the mind bears less resemblance to a fact-gathering or puzzle solving apparatus than to a population of foraging organisms. The foraging entities are interests that grow from rewards, just as populations of organisms grow on the basis of food sources. Like the variety of organisms that stably occupy a habitat, interests fill various motivational niches on the basis of their specific abilities to maintain and defend themselves there, which largely depend on the time courses of the rewards on which they are based."

1.2 The role of mathematical approaches

Our belief, which we shall argue in detail in subsequent chapters, is that such traditional metaphors as "dynamical system", "machine", or "computer", which have underlain the application of mathematics to other scientific domains, are inadequate to encompass the essential characteristics of adaptive autonomous systems. In particular, it will be essential for the theory, whatever form it may take, to encompass "intentional" (or, "subjectivist" vs. "objectivist", or system-centered vs. observer-centered) perspectives. This will hold even for the case of such "simple" autonomous systems as *C. elegans* , to which one surely would not attribute "volition", let alone "higher cognitive function". The attempt to construct a genuinely scientific theory on such a "subjectivist" base is a radical step, and many would argue that it is ipso facto self-defeating, if not self-contradictory. We believe, on the contrary, that what constitutes "our" science in its present form (and, similarly, "our" mathematics) is, to a considerable extent, conditioned by historic contingencies (in particular, the types of natural phenomena chosen for investigation), and should not be viewed as normative of "any possible" science.

In particular, it is not clear to us, for example, that the general idea of "mathematization" of science which was so successful in physics and, in principle at least, in chemistry, can be applied, even in principle, to an autonomous system context. In other words, the traditional ideology that "mathematical physics" is the ultimate state of an "exact science", may be totally invalid in the present context. The whole methodology of physics, engineering science, and modern technology in general, is the modeling of the systems one wants to describe, in a very precise terminology, either using infinitesimal calculus or other kinds of "calculus" (symbol manipulation), in such a way that one can simulate the systems, predict their behavior, and act on them, if necessary.[5] In the setting of autonomous systems, the role of theory may be rather different, and less centered on sharp prediction of system behavior. We anticipate that this role will be to provide a framework for discussing the nature of the "coherence" of the system, and how it is achieved and maintained. Typically, whether in the context of natural or of engineered systems, the issue of "coherence" is treated as, at best, a background consideration, and certainly not as a focus of inquiry in itself. We believe that any theory of autonomous systems will need to bring "coherence" into the foreground. We expect that, correspondingly, new kinds of mathematics will be necessary.

It may well be that there are "laws", or principles, of coherent organization. However, it is doubtful that these take on the character of physical laws, such as Newton's laws or even the laws of quantum mechanics. Considered in an engineering

---

5 As one moves away from the traditional "hard core" sciences towards, say, the cognitive sciences it is common to encounter such arguments as to what kinds of theory are required. A notable example is the controversy concerning "strong" artificial intelligence, or "computer functionalism". (See, e.g., Simon, Newell, and Vera as proponents, and H. Dreyfus and J. Searle for two types of dissenting view, among many possible references.) Nor is it a coincidence that such debates take place also in linguistics ("two" of the viewpoints put forward being those of "generative grammar" and those of "cognitive grammar").This question arises also in logics (beginning at least as early as the twelfth century with the "quarrel" between the Dominicans and Franciscans over universals and the question of nominalism, continuing later with Descartes' dualism, Leibnitz' monads, Husserl's (and subsequently Heidegger's) phenomenology, and into our own day under various names).

context, they are much more likely to be akin to the "principles", explicit or intuitive, used by a creative designer. A case in point is the design of "layered hierarchical systems" (cf. Section 2.6e)). Physical laws have a role to play here, but the final design incorporates principles which may not be expressible within current mathematical frameworks.

1.3 Organization of the paper

In Chapter 2 we present some examples of complex interactive systems, both artificial and natural, and examine a variety of approaches that have been applied to the study of these systems. Our aim here is two-fold: to highlight tools or viewpoints that we believe can be transposed to the setting of adaptive autonomous systems and, conversely, to clarify why certain other approaches are ill-suited or inadequate for this setting.

In Chapter 3 we attempt to isolate some of the basic characteristics of adaptive autonomous systems. In addition we discuss some elements that we feel must enter into a theory of such systems, as well as the kinds of role we envision for mathematical approaches.

In Chapter 4 we contrast the notions of "constraint" and "specification", in conjunction with the two dichotomies "natural/artificial" and "embedded/simulated". This leads into a discussion of distinct types of "failure modes" and "evaluation criteria" for the distinct system types.

In Chapter 5 we analyze the notion of "behavior", with an emphasis on its emergent character, and elaborate on the central notion of the "value system".

In Chapter 6 we enter into a more detailed discussion of "control". This is facilitated via the introduction of the useful auxiliary notions of "state" and "module".

A recurring theme in the preceding chapters is the rationale for the emergence of a process of "symbolization" (i.e., the construction and manipulation of symbols), in a sense quite different from the usual computer-science use of the term, and largely non-computational in character. We refer to this process as "neural symbolization" since our thinking here is primarily influenced by, and directed towards, biological organisms endowed with nervous systems. In Chapter 7 we develop this theme in some detail, and put forward some elements of a "neural symbol"-based schema for an integrative functional architecture. This schema, which is closely tied to the value system, incorporates the formation of associations, categorization, and planning and decision-making.

We wish to emphasize the fundamentally exploratory and provisional character of the present study. Our aim here has been relatively modest: to begin the process of intellectual reorientation which we believe necessary for the study of adaptive autonomous systems. In fact, much of our effort is directed at the preliminary task of identifying, and giving definition to, a subject area. Thus, rather than carrying out a sharply focused and systematic investigation, we approach the subject from a number of directions, and broach a range of topics. In addition, our treatment is still at a pre-technical level. This leads to a problem of terminology. We expect that, ultimately, the subject will require a technical vocabulary of its own. However, given our current state of understanding, we have felt it more appropriate for the present purposes to borrow terminology from the various more established disciplines which impinge on the subject.

When transposed to our setting, this borrowed technical vocabulary should be construed only in a quasi-technical sense, as "pointers" to as yet ill-defined notions. In particular, we often intend to draw upon only certain of the various connotations (sometimes coming from distinct disciplines) attached to a particular term.

The above caveats notwithstanding, we believe that we have succeeded in bringing to light and weaving together a number of important themes (among them "coherence", "internalist perspectives", "emergence", the "value system" (basic, higher-order, and symbolic), "vague specifications", "non-algorithmicity", "neural symbolization"), which recur repeatedly throughout the paper. We feel that the work presented here can serve as a basis for more systematic studies, which we plan to undertake elsewhere. This includes, in particular, the development and elaboration of more comprehensive schemata for integrative architectures.

2. Various examples of complex systems and traditional approaches

A. Complex interactive systems

2.1 Computer-like systems

The abstract notion of a computer-like system is the notion of automata or, more generally, of a Turing machine, which is supposed to be the most general manipulator of symbols or the most general computing-like device. These systems are characterized by a transformation input $\rightarrow$ output working according to prespecified mechanisms. These mechanisms are fixed once for all, and can only be changed by an external intervention. In particular, these machines do not adapt and do not learn.[1] The machines do not have any environment, in the sense that they do not have to act on their environment and get reaction from the environment. In particular, time is not important (it matters only as the "length" of a computation).

2.2 Various networks, manufacturing control process,....

Again these systems receive various inputs which should lead to desired output (according to some criteria of evaluation which have to be optimized).

The difference with computer-like systems is that these systems do work in a given environment, and are subject to feedback (or purely reactive) control which adjusts the functioning of the system when the desired output is not obtained. These systems can tolerate a certain amount of error. Time becomes important and a real factor because of synchronization questions, and because environmental requests have to be fulfilled without too much delay. Still, these systems are not adaptive. They cannot modify "their" rules or evaluation criteria, and they do not learn.

2.3 Physical systems

These are structural systems where behavior can be determined and predicted successfully from the "laws" which govern it, in conjunction with mathematical reasoning. These laws are conservation laws and optimization laws.

2.4 Biological systems, decision systems

These systems seem to be of a rather different class than the previous systems. They in general do not associate, in any precise way, a given output to a given input. Rather, they have to maintain certain relations with their environment, and these relations are partly fixed by the system itself.[2] In particular, these systems can adapt themselves to certain changes of the environment. It is clear that they can only adapt to certain classes of environments, and if the environment fluctuates too wildly these systems can become disorganized or "disappear". Needless to say, time is a very important variable for these

---

1 We do not consider "machine-learning" programs as genuine counterexamples, on several counts.

2 Indeed, in various senses, they themselves determine which aspects of their embedding milieu actually count as their "environment. (See, for example, R. Lewontin, " The organism as the subject and object of evolution", Scientia 118, 1983, pp. 65-83).

systems. Biological systems function, in particular, as control systems. However, they incorporate principles of adaptation and learning which go well beyond those supported by the control systems which are currently available for engineering applications. That is, control theory as currently constituted (i.e., feedback theory for linear and non-linear multi-variable systems, deterministic and stochastic optimal control, theories of robust control) cannot accommodate the adaptive and learning behavior manifested by biological systems.(We shall not discuss here economical systems or ecological systems which are, in some sense, related to biological systems, as noted in Chapter 1).

2.5 Other distinctions : natural vs. artificial, embedded vs. simulated

These are other very important distinctions which are "transversal" to the previous distinctions

a) Natural / artificial.

This is a very important distinction : On the one hand we have natural systems like a cell, a group of cells, an immunological system, a large social organism, physical or chemical systems, etc. On the other hand we have artificial systems like robots, computers, networks, machines, manufacturing control processes,... It is clear that artificial systems which are actually implemented (and not only simulated) rely on natural systems for their functioning (like motor for the robot, electronics for the computer) and, as usual, the distinction between natural and artificial is not completely sharp.

b) Intentional / non-intentional

Artificial systems are non intentional in an intrinsic sense. Surely, they serve purposes, but these purposes are global purposes imposed by their designers. On the other hand, certain natural systems have an intrinsic(vs. extrinsically imposed) intentionality. They have their own purposes, which they can modify. They can form new purposes, and in fact they often do so. They exhibit intrinsically derived (and, in some cases, volitional) behavior. In particular, we argue that intentionality need not be simply a viewpoint or "stance" which an external observer can make use of in order to provide a rationale for, or to predict, the behavior of the system.

Again the distinction is not completely sharp: Does a (living) plant have purposes? Do immunological systems ?.... In any case, one of the features of an autonomous system is the ability of the system intrinsically to form its own purposes, modify them, and select some of them, at a given time, from among a certain range of possibilities.

c) Embedded / simulated.

An easier distinction is the distinction between an actually built and embedded system, and a system which is purely formal, or exists merely as a simulation (in general, on a computer). This distinction is also very important because autonomous systems are always embedded, and have to deal with an actual environment. The essential difference lies in the fact that embedded systems are subject to material implication, and not simply to formal implication. In recent years there has been debate within the field of AI (artificial intelligence) on the issue of "situatedness", two foci of the debate being implicit

vs. explicit modes of "knowledge representation", and the importance of "context".[3] This point of view has also led to work emphasizing actual physical embeddedness in the world.

In connection with autonomous systems it is useful to make a further (at this point, intuitive) distinction, between "embodied" vs. "embedded" systems. An embodied system is embedded, but in addition, has the following characteristics: (i) The system is an "organism", with an "internal" or "subjective" viewpoint. (We are not thinking here only of organisms with awareness, let alone self-awareness). (ii) The system's body serves as a "ground reference" for the system to construct this viewpoint. It provides the system with the means to acquire (or bootstrap) both a kind of "semantics", and a kind of "pragmatics", the latter via a bodily-mediated "value system". (We shall discuss the notion of value system in some detail in Chapter 5). Thus, both "meaning" and "value" are emergent, rather than externally imposed. (The sense of "self", in those organisms which possess it, would be emergent in the same sense).[4]

2.6 A range of metaphors

We can now introduce a rather different category of distinctions between various systems, linked to the types of analyses, or metaphorical descriptions, typically made of these systems. Generally, a given system cannot be fully characterized by just one of these "metaphors" , but in practice this is neglected, and usually one attempts to describe the main features of certain systems via a single dominant metaphor.

a)The mathematical physics metaphor

This serves mainly to describe and model, but also to control, physical and chemical systems and, in technology, all systems (provided they are not too complicated). For modeling, this approach relies on the mathematical description of physical laws (the language of ordinary, functional, or partial differential equations; in particular, dynamical systems). The control of such systems can, in principle, be accomplished using the theory of control for differential equations. A good example on the modeling side is celestial mechanics. A counterpart on the control side is the control of a spacecraft or the attitude control of a satellite. The approach of mathematical physics is, in some sense, the oldest and most successful method of analysis, and when it is applicable it gives, in general, extremely good predictive results.

b) The computational metaphor

This metaphor describes the realm of computing system devices, both software and hardware, from the standpoint of abstract functionality (or logical implication) vs. physical embodiment (or material implication). Its basis is the Turing machine, or its

---

3 See, for example, B.C. Smith, "The owl and the electric encyclopedia", Artificial Intelligence 47 (1991), pp. 251-288, and references cited there. In this very interesting article Smith discusses foundational issues and guidelines for a theory (and practice) of "embedded computation", which he links to the notion of "partially disconnected participation". (He points to "self-involvement" as being more relevant in this setting than "self-reference".)

4 In the roles we attribute to embodiment, we are close to the views of A. Damasio. (See "Descartes's error", forthcoming, 1994, Putnam, Grosset Books). See also the references in the footnote to Section 7.12c)(iv).

various modifications, such as the RAM (random access machine), and symbol manipulation. Symbols are manipulated abstractly as independent entities, that is independent of the environment, of the method used to manipulate them, and of the content or meaning they are intended to convey. It is clear again that this method of analysis is extremely successful when it is applicable, namely in the design of software, and in the analysis of simple information systems.

It is also clear that, in contrast to the mathematical physics metaphor, it is not usually applicable by itself, in the sense that software requires physically embedded hardware in order to run; even the symbols need to be physically instantiated. The description and analysis of the hardware, as a physically embedded system, will be ultimately done using the mathematical physics metaphor.

In both cases, the mathematical physics metaphor and the computational metaphor participate to the general ideal of an exact mathematical description and analysis, and they rely on the traditional algorithmic mathematical ontology. It is also clear that these two methods of description, when they can be applied, can enjoy a complete success. Unfortunately, they are somewhat brittle, and when they are inapplicable or inadequate, their failure can be almost total.

c) The machine metaphor

By "machine" we refer to a physically embedded system, with which one associates some specific functionalities or purposes. Thus, it makes sense to speak of a machine as "malfunctioning" or being "broken", whereas it would not make sense, in general, to speak of a physical system in these terms. In many cases the machine may be treated via the mathematical physics metaphor. One the other hand, it may also be treated somewhat non-mathematically, as a "mechanism". By this we mean an interconnected nexus of physical cause-effect relations.

d) The control metaphor

This arises in the engineering setting, in at least two contexts:(i) Design of a system to perform certain functions, subject to a range of guidelines, specifications, and restrictions. (ii) Maintenance of a system's behavior within specified operational guidelines. Here one begins with a system that is largely already designed as a physical plant, for which one provides a "controller". Whether the controller is realized as hardware or software it will likely need to function in "real time". From the standpoint of autonomous systems the control metaphor has the following drawbacks. The metaphor suggests, which is in fact the case in practice, that "control" is imposed from without, even if it is realized "internally" within the system. Similarly, the designer, rather than the system itself, determines the "problems" the system is to solve. In particular, these problems do not emerge from the internal "needs" or purposes of the system.[5] This emphasis on the imposition vs. emergence of control can be misleading as to the nature of the latitude available in the overall, integrative system design. Thus, while it may take into account design constraints (i.e., constraints as to which design specifications are

---

5 This is not to say that the designer's purposes are totally orthogonal to system-centered purposes. For example, the designer tries to assure that the system does not break down or lose its integrity as a system. However, from the design standpoint this consideration is secondary, whereas from the system standpoint this is the base from which all system activities arise.

permissible) due to the physical embedding of the system, it does not take account design constraints imposed by the requirements of "coherence" of the system's overall functioning and behavior. That is, it does not take adequate account of "compatibility" conditions among design specifications (or, alternately phrased, global constraints on the "gluing together" of local controls).[6] Curiously, this presupposition of design latitude results in an actual "brittleness" of design. The old "cybernetics" picture does convey a sense of integrative organization, and an implicit emphasis on "coherence". However, it does not make this notion explicit and, in effect, provides only a single mechanism for achieving and maintaining such coherence, namely "feedback". It is far from clear that this is adequate. In particular, cybernetics, with its emphasis on feedback, focuses on the maintenance of *static* equilibrium (steady-state as well as fixed-point), both with the external environment and within the internal milieu of the system. For an adaptive autonomous system, on the other hand, coherence is linked to *dynamic* equilibrium.[7]

e) The network theory metaphor and layered hierarchical systems[8]

We shall discuss the metaphor of electrical network theory in conjunction with "layered hierarchical designs", such as the design of a communication network permitting the transmission of data, voice, and images.

Network theory deals with the relations between currents and voltages which express the laws of behavior of electrical networks. In the most elementary setting these relations are the Kirchhoff laws. It is significant that the cause-effect relation is not made explicit a priori, but is instead left as a choice of representation, to be decided according to the purposes of analysis. The relationship of the network to its environment is realized via the "load" on the network. The network laws describe the equilibrium (coherence) that is maintained between the network (the system) and the load (environment). Networks can be combined by interconnecting them in cascade, in parallel, or in feedback. The laws of the resulting network are derivable from the laws of the subnetworks which have been combined and the interconnection constraints.

The design of modern engineering systems such as communication networks incorporates a "layered hierarchical structure" in the following sense (not to be confused with standard notions of "hierarchical" or "tree-like" organization). Here, the level of "abstraction" increases as one moves up the hierarchy, with the lowest layers being the least abstract. Typically, the lowest layer is the "physical" layer, consisting of interconnections of various subsystems. These subsystems are modeled (approximately) using physical laws (viewed as relations between physical variables), and the subsystem models are then interconnected as in the interconnection of electrical networks. The signals involved are physical signals (current, voltage, position, velocity), modeled as analog signals (in the time or frequency domain). Each higher layer takes an abstract view of the preceding lower layer. In a communication network the next layer above the

---

6 This issue is emphasized in the *C. elegans* Working Paper referenced in Chapter 1.

7 In using the term "dynamic equilibrium" we mean, in particular, to take account of transitions between distinct static equilibria, including the formation of new equilibria. However, we do not wish to limit ourselves to a (parametrized) dynamical system framework (cf. Section 4.4c).

8 A discussion and analysis of the network metaphor may be found in J.C. Willems, "Models for dynamics", Dynamics Reported, Vol. 2, 1989, pp. 171-269. (See also the reference to the work of Willems in Section 2.8.) Layered hierarchical structures are discussed in D. P. Bertsekas and R. Gallager, "Data networks", Second Edition, Prentice Hall, New Jersey, 1992.

physical layer is the protocol layer (protocol = distributed algorithm), where the signals are bit strings which are manipulated using discrete (combinatorial) algorithms. The two layers in this instance have to be "interconnected" by properly designed interfaces which allow the transformation of signals (analog) to symbols (bit strings) and vice versa. It is a principle of good engineering design that if the layers themselves have been designed properly then the design of the interfaces will be simple. Nevertheless, error detection and correction mechanisms have to be put in place so that the layered hierarchical system functions in a coherent manner.

As emphasized above, the design principles involved here, unlike those of structured programming, do not lead to a tree-structure. Each layer is "homogeneous" and robust, in the sense that uncertainties arising in a particular layer are neutralized as much as possible within that layer itself, and not allowed to propagate to a higher layer. Moreover, mechanisms of adaptation and learning are "built in" within each layer in order to deal with parameter variation and, to some extent, changes in the environment. In current designs, the "abstraction" is carried out according to principles which are independent of context, and the symbols at higher levels are genuinely abstract, being themselves devoid of content and meaning. Even for engineered systems, such as the above layered structures, one needs a better understanding of the potential role of context (and of semantics), and how it might be incorporated into the design.

f) The information processing system metaphor

"Information processing" is used in at least two distinct senses, but with the distinction often being elided in practice. The first is information in the sense of the Shannon's theory of a communication channel. Here "information" is not linked to meaning or content, but only plays the role of disambiguation. The second approach to information processing systems emphasizes "information" as conveying actual meaning, but with the information, and associated meaning, being independent of the "apparatus" which produces or manipulates this information. As examples, we can cite various communicating devices where the content of information becomes important (or at least certain aspects of the content, and where the usual abstract computational approach will not be successful. This includes systems where one cannot be content with traditional time-independent and context-independent logic (such as the first-order predicate calculus), but where one wants to go beyond that, using various non-standard logics like temporal, non-monotonic, modal logics, etc. The information processing metaphor in this "semantic" sense is generally used informally as, for example, in discussions of nervous system function. However, there has also been work on formal theories of semantic information, for example in connection with formal treatments of linguistics, an "early" example being Montague grammar.

g) "The" biological metaphor

It would be misleading to speak of a single "dominant" biological metaphor. Certainly, the field of biology has given rise to the fundamentally new metaphor of "natural selection". However, as regards functional organization, the field has largely drawn its metaphors from outside, rather than generate its own[9], the primary ones being

9 We do not regard "neural networks" as an exception.

those noted above. Our point of view, which we shall seek to explain and justify in the remainder of this paper, is that, as regards integrative organization in a setting of autonomy and adaptiveness, these metaphors will not suffice. The present work is one attempt to go beyond them.

Our reservations as to the adequacy of the traditional metaphors extend even to the context of such "simple systems" as *C. elegans* . One may ask with some justification, what else is *C. elegans* but a tiny machine, or automaton, albeit the product of natural evolution vs. a human designer's blueprint ? What other kind of description is even conceivable ? This, it is fair to say, represents the consensus viewpoint of the *C. elegans* research community towards their model organism.[10] In our opinion, this viewpoint is in error. Certainly, to gain an understanding of *C. elegans* from a mechanistic perspective is an immensely challenging problem. It is an integral, if not essential concomitant of any attempt to discover, and give expression to, the organism's principles of organization. At this point one can only speculate at what these principles may be, or even what kinds of thing they may be. However, we believe that any attempt to express these principles in exclusively mechanistic terms is doomed to failure, that it is in principle (no pun intended) impossible. Not only is the mechanistic perspective "incomplete", in the sense that there are interesting and important non-mechanistic properties of the organism , but it is not "self-contained"; That is, any attempt to account even for the mechanistic properties of the organism will need make reference to terms not themselves expressible within the mechanistic frame.[11]

B. Some approaches to complex interactive systems: engineered systems

In this section we briefly touch upon aspects of several approaches which have been applied to the description, design, and analysis of complex interactive systems in a computer science or engineering setting. Biologically oriented approaches will be touched upon in the next section. Our purpose is to clarify the extent to which the approaches discussed in the present section can or cannot usefully be transposed to the setting of adaptive autonomous systems. Our presentation will, for the most part, take on

---

10 Paradoxically, it is the common biological provenance of our object of study and ourselves which makes it seem to us like the artefact of an alien technology. With hindsight this alienness, or "illogical" (in particular, far from cleanly modular) organization (to use Brenner's term; cf., Brenner: Why Is Development So Illogical), is only to be expected in any biological organism, even one as "simple" as *C. elegans* . In fact, it has, until comparatively recently, been a widely held view among neurobiologists, both in and out of the field of invertebrate neurobiology, that invertebrates (as contrasted with vertebrates, especially mammals or primates) are "simple systems". That is, not only are they machines, but automata of comparatively straightforward "logical" design, that would make good sense to a human engineer. Gradually, this viewpoint is becoming displaced by a recognition that these "simple systems" are far from simple in the above sense. (Cf. Harris-Warrick et. al. ; Mpitsos and Soinila.)

11 We categorically deny that "non-mechanistic" implies a recourse to vitalism.

In some sense *C. elegans* is on the "boundary" between machine and non-machine and hence is particularly apposite for studying where the machine metaphor reaches the limits of its validity, and precisely how it becomes inadequate. In fact, one does not need to go all the way to a metazoan with a nervous system, like *C. elegans* , in order to reach the "boundary" between machine and non-machine. A case (of a different nature, and admittedly more difficult to argue) can be made to this effect for a single-cell organism such as *E. coli* . An interesting research program on *E. coli* as an integrated system is discussed in (Mittenthal and Levinthal).

a negative cast. In large part, this is because we feel it will be most valuable to use these approaches as a foil, to bring into relief essential distinguishing characteristics of adaptive autonomous systems. Our remarks should not be construed as criticisms of these approaches per se, since we are, after all, seeking to transpose them to settings for which they were not initially intended. Our criticisms in this section also are, by and large, not directed to these approaches in their potential role as "meta-level" tools, in conjunction with an integrative modeling framework. That is, these approaches may well be useful in conjunction with the reasoning *we* carry out about the system, even though they may bear little or no relation to the processes (cognitive or otherwise) that the system itself carries out. [12]

2.7 Approaches to be considered

We begin with a listing of the approaches we will be considering.[13] The major headings are listed, roughly speaking, in increasing order of concreteness. The ordering of the subheadings is more idiosyncratic and, indeed, there is significant overlap between various of the "distinct" approaches, including the major headings. This listing is for convenience of reference. We shall make no attempt at an actual review or survey of these approaches, even in caricatural form, nor will we even discuss them one by one.

1.Theoretical computer science and applied logic
- 1.1 Formal specification and verification methods
- 1.2 Non-standard logics (temporal logics, modal logic, non-monotonic logic, fuzzy logic, etc.)
- 1.3 Concurrent process theory (Hoare, Milner)
- 1.4 Statecharts (Harel)
- 1.5 Semantics of programs (denotational semantics)

2. "Abstract" control theory
- 2.1 "Traditional" control theory (classical feedback control, time-domain state-space models/optimization of control laws)
- 2.2 Discrete-event control (Wonham-Ramadge-Ostroff)
- 2.3 Hybrid control

---

12 Such an integrative modeling framework is a component of the *C. elegans* research program mentioned in Chapter 1. The modeling framework involves the construction of a network of mathematical/computer "partial models" concurrently representing, from multiple experimental and analytical perspectives various facets, components, and organizational "levels" of the organism. These partial models and the constraints linking them are to represent our "picture" of the organism, rather than the organization of the organism itself. Computer science approaches such as those discussed in this section (as well as certain AI approaches) may well be appropriate tools for carrying out manipulations and analyses of this network of models, which is a formal mathematical entity, albeit of nonstandard type. In the setting of this network, "clean", formal, "externalist" (see Section 2.8) approaches, emphasizing functoriality, may be precisely what is needed. One example of a potentially useful tool is the "concurrent constraint programming" approach of Saraswat, which emphasizes partial information and entailment among partial information (see, e.g., V. A. Saraswat, "Concurrent constraint programming", MIT Press 1992). We would need this in a form incorporating non-monotonicity.

13 This list is not entirely idiosyncratic. It includes approaches we experimented with in the earlier phases of the present work.

3. Genetic algorithms (Holland, Koza)

4. Integrative robotics architectures
    4.1 Subsumption architecture (Brooks)
    4.2 Computational neuroethology (Beer)

There are basic obstacles to transposing the above approaches to the setting of adaptive autonomous systems. These obstacles are not associated simply with technical difficulties, but are deeply rooted in the very nature of the approaches.

2.8 Internalist vs. externalist perspectives

There is no reason that theoretical treatments of autonomous systems should avoid externalist (or "outside-in") perspectives, i.e., viewpoints taken by the designer (in the case of an engineered system) or by an observer of the system. However, it is internalist perspectives (i.e., those arising from the system itself) that must serve as the foundation of the theory. These internalist perspectives are lacking in the above approaches.[14]

For example, the thrust of "formal" approaches (such as 1.1, 1.3, 1.5) is towards "invariance"; i.e., towards implementation-independence (both hardware and software), context-independence, etc. In this setting, any particular realization is intended as an instantiation of the same pre-existent abstraction. In the autonomous system setting, on the other hand, everything is highly dependent on the individual system, on its history, and on the context. The system organization is "emergent" rather than being an "instantiation" of anything external.

An additional example of this externalist perspective is the philosophy of Olderog and Hoare for modeling of processes.[15] According to this viewpoint a process is equated with the set of "observations" which can be made of it. A similar point of view, in the setting of control theory, has been put forward by J. Willems, in his behavior-based approach to control of dynamical systems.[16]

One concomitant of internalist perspectives it that one can no longer rely on external a priori "yardsticks" (such as the universal computability of Turing machines or the universal rationality of game theory; see Section 3.7a)) as "ideal" standards against which to measure the actual system. However, we do not regard this as a weakness of the internalist perspectives, but rather an indication of the artificiality of such "ideal" yardsticks (cf. Section 7.12, including footnotes, for some of the difficulties associated with ideal rationality).[17]

---

14 For another perspective on internalist perspectives see N. Chomsky, "Language from an internalist perspective", preprint, 1991.

15 See, for example, the discussion and references in V.A. Saraswat, M. Rinard, and P. Panangaden, "Semantic foundations of concurrent constraint programming", (POPL 91).

16 See, for example, J. C. Willems, "Paradigms and puzzles in the theory of dynamical systems", IEEE Transactions on Automatic Control, Vol. 36, No.3, 1991, pp. 259-295. In the "behavior-based" approach, the "state-space" is secondary and non-intrinsic, and depends on the model representation. The same is true of "inputs" and "outputs". What is intrinsic is, for example, "controllability", which reflects the capacity to modify system behavior. (An illustration of the construction of a state-space realization from the space of behaviors, is the Nerode equivalence of finite automata theory).

17 We have the impression that an insistence on externalist perspectives, together with the corresponding "ideal" yardsticks is closely related to the "foundational" paradoxes. We touch upon this in Section 7.3, but hope to examine the question more thoroughly elsewhere. Similarly we hope to make at

2.9 Symbols as emergent vs. imposed

Another manifestation of the outside-in perspective, linked to the emphasis on formal vs. embedded systems, is the traditional distinction between syntax and semantics, in particular the idea of a functorial syntax ---> semantics mapping. In particular, in these formal approaches a symbol, as well as the correspondence of the symbol to its referent, can be introduced arbitrarily. The symbol need not emerge from within the system, and its meaning need not be tied to its physical action within the system. In biological systems, on the contrary, symbols really do something, and thereby acquire their symbolic function.

This formal character figures also in the discrete-event and hybrid control approaches (2.2 and 2.3). For example, in the case of analogue/discrete systems, the analogue character tends, in practice, to be introduced by the modeler, as a particularly convenient or efficient mode for working out a pre-determined abstract computational problem. Thus, the analogue processes are supervenient on the abstract computation, rather than vice-versa.

2.10 Emergence of behaviors

A fundamental problem concerns the notion of "behavior" of the system, and how these behaviors are controlled or coordinated. This is linked to the notion of "state". Typically, in the approaches listed under 1 (and also, to a large extent, 2), a "behavior" is a sequence of elementary actions, where an "action" may, for example, be represented as a transition operation on a given "state space". Our concern here is not with the question of whether "behaviors" or "states" have primality, but with three more basic issues:(i) The first is the "extensionality" of the treatment. We are uncomfortable with the notion of a "behavior space" as some kind of Platonic, pre-existent set of potentialities.[18] (ii) The second is the existence of basic building blocks, in this case the elementary actions. We think that this reductionist type of viewpoint does not do justice to the context-dependence and viewpoint-dependence that will be required. (iii) The third concerns sharpness of characterization. We think that a certain degree of "blurring", or "vagueness" is intrinsic to the autonomous system setting, which must be accommodated by the formalism. Nor can this be done, we believe, by starting with a sharp notion and factoring out by an equivalence relation. For example, the notion of "behavioral equivalence" in Milner's CCS calculus is too sharply drawn for this setting. Similarly for notions of "partial information", as in D. Scott's denotational semantics. A "partial specification" of an entity suggests a pre-existent feature-template, with certain slots filled in and others left blank; a "vague specification", on the other hand, would not presuppose such a pre-existent template. (We do not mean to deny the existence of relevant constellations of features. However, we expect that these, in general, will be emergent and context-dependent. Moreover, the potential "contexts" cannot be

---

least some preliminary comments on the questions associated to "intersubjective agreement" (or intersubjective "comparison" or "communicability") raised by internalist perspectives. This issue of "external yardsticks", we feel, also bears upon the topic of "directionality" (towards increased "perfection" or "complexity) of evolutionary processes.

18 At the very least, we do not see (in a discrete setting) why the system's behaviors should be recursively enumerable.

enumerated in advance. In particular, we strongly doubt that "context" can be accommodated simply as an additional slot or list of slots in a predetermined feature-template.) We do not think that "fuzzy logic" provides the solution since, in practice, it begins with a sharp picture which is only subsequently "fuzzified", and then in a rather precise way. This problem of pre-existent templates also pertains to "possible world" semantics and the associated modal logics.[19]

We are somewhat happier with the treatment of "behaviors" in the approaches listed under 4, but still have basic concerns here. First of all, in these approaches behaviors are forced on the system. For example, in Beer's simulated "artificial insect", one "wires" in "edge-following behavior". We would expect, on the contrary, that most complex behaviors are really emergent properties of the system (on a somatic time scale) and are not "pre-wired".[20] A related concern we have is that the issue of genuine "behavioral choice" from among "competing behaviors" is not considered. Hierarchies of behaviors (with some provision for a "reactive" form of context-dependence) are fixed once and for all, and conflicts are pre-resolved in the specifications of the system.[21]

2.11 Combinatorial explosion of the state space

The same objections carry over to the treatments of "state". In particular, we do not think that approaches listed under 1 and 2 will be helpful in dealing with such notions as "internal state" of the organism or "state of alertness or arousal" of the system, let alone such notions as "cognitive state" (for example, as in Chomsky's picture of "language acquisition").

In addition, the same considerations give rise to the risk of combinatorial explosion, and to the question of control of "search". Granted, approaches to process theory tend to circumvent the use of a "state" notion precisely so as to evade this combinatorial explosion problem[22] but, at least in our context, we feel this only hides rather than eliminates the problem.

---

19 This is likely to be a difficulty with any form of "objectivist semantics". (See the reference to the work of M. Johnson and G. Lakoff in the footnote to Section 7.12(iv)).

20 See, for example, R.D. Beer, "Intelligence as adaptive behavior", Academic Press, 1990. We do not deny that some behaviors of biological organisms (in particular invertebrates) are, to a greater or lesser extent, pre-wired.

21 Some work has been done to incorporate "learning" into the subsumption architecture framework. (See, for example, R. A. Brooks, "Intelligence without reason", pp.569-595, IJCAI-91).

22 A classic illustration of this is G. Kahn's "determinate data-flow" model of determinate concurrent computation. (See G. Kahn, "The semantics of a simple language for parallel programming", in J.L. Rosenfeld, ed., Proceedings of IFIP Congress 74, pp. 471-475, Aug. 1974).

As an illustration of the combinatorial explosion problems that can arise: We experimented with a state-space formulation of a variant of R. Milner's calculus, which we applied to the canonical examples in his book ("Communication and concurrency", Prentice Hall, 1989), notably the "scheduler" example. We immediately encountered a combinatorial explosion of the number of states, formulas...

We have also experimented a bit with the statechart approach of Harel, which is intended to mitigate combinatorial explosion via a variety of devices, including a hierarchical clustering of states. We have found this clustering somewhat unwieldy, leading to a confusion between the state of the system itself and various parts of the system which are described at the same level. In addition, the statechart approach may be inherently too flexible to provide, of itself, adequate control of complexity. For example, it permits arbitrary C-code, unrestricted in complexity, to be associated to the node of a graph.

2.12 Control and coordination of behaviors

In the approaches listed under 1, the question of control of behavior is not explicitly raised (except, perhaps, in connection with control of search). In the approach 2.2, the issue of supervisory control of state transitions is explicitly introduced at the foundational level, but is limited to the disallowing of certain classes of state-transition. This, together with its restriction to a finite-state automaton (or regular language) framework, limits the relevance of the approach for our context.

In addition, as discussed in Section 2.9 above, neither of the approaches 2.2 or 2.3 works via embedded vs. formal symbols. Just what a difference this makes can be appreciated by considering the regulatory and control processes within a single cell, e.g. the processes of gene regulation or cell metabolism. One might expect here that a picture of discrete-event control would be highly apposite, but the formal symbol character of the approach surely loses much of the essence of the underlying processes (e.g., the way in which feedback inhibition, say, is actually mediated; the way in which regulatory pathways interact, etc.). This formal symbol character also overemphasizes the distinction between controlling and being controlled.

A significant problem with the approaches 2.1 and 3 is the reliance on optimality criteria. We are not concerned here with the question of how difficult it may be for the system (or the designer) to actually compute the global minimum (and, in particular avoid becoming "trapped" in a local minimum). Our concern is with the more basic issue that, at least in the case of an organism, there is in general no canonical optimality criterion, and the underlying control organization does not rest on optimization. At a minimum (pun intended) there are trade-offs and choices to be made among a range of value criteria.[23] This raises also the issues of behavioral choice and conflict resolution mentioned at the end of Section 2.10.

2.13 Motivation for action

In the approaches listed above there is no intrinsic relation of the system to its environment (with the term "environment" suitably interpreted for the various respective approaches). There is no system-centered vs. designer-imposed reason for the system to do anything in particular, or anything at all., i.e., either to modify its internal "state" or to interact with its environment. By "reason for action" we mean system-internal needs or system-centered and system-derived goals whose satisfaction requires the system to draw information or resources from the environment, or to act upon the environment or itself, or to refrain from certain actions. That is, (as we shall discuss more fully in Section 5B), a system-centered framework of "values", so that actions or inaction by the system are "relevant" to and carry "consequences" for the system itself (with relevance and consequence "measured against" the value-framework, which is itself grounded in the system). By "acting upon itself" we include modifying its structure, its "rules" of behavior, its goals, or even (to some extent) its framework of values. We believe that genuine "learning" or "adaptation" must be linked to this kind of system-relevance, as determined via the system's own value framework. We do not consider the pre-wired

---

23 Even in the setting of natural selection, one cannot really regard "fitness" as a utility-function to be maximized. If it were, relative fitness would be a transitive relation.

behavior hierarchies of the approaches listed under 4 as system-centered in the above sense.[24]

The above-noted non-intrinsicness in the system/environment relation associated with approaches 1-4 is closely linked to the fact that these approaches do not take into account inherent constraints on the system.[25] The main constraints arise from the fact that the system has an environment and has internal needs. The environment fixes the three basic constraints : spatial constraints, chronological constraints and resources constraints (or space, time and matter-energy). The fourth basic constraint comes from the system itself, namely the system needs resources to survive and one of the basic things it has to do is to maintain its metabolism. These constraints are tied to the embeddedness and, to an extent, the embodiment of the system (see Section 2.5).

A related issue is that the above approaches do not lay stress on the fact that the existence of the system, and its integrity as a system, cannot be taken for granted. A central task of an autonomous system is to maintain itself as a system.

2.14 Coherence

We have already raised the problem of "coherence" of behavior and control organization in Section 2.6d). None of the above approaches addresses this problem directly, or as a central concern. For example, the integrative robotics approaches listed in 4 take a largely ad hoc (based on neuroethological empirics, design experience and intuition, etc.) approach to the question of which behaviors can (or must) be included, and how these behaviors must be coordinated, in order to achieve and maintain coherence. In

---

24 Beer does, in fact, endow his simulated artificial insect with a form of "internal motivation", via a feeding arousal system linked to sensing of an internal energy level which runs down unless replenished via "eating". In our terminology in Section 5B this, in a sense, corresponds to a "primary value". However, it is problematic whether we actually wish to attribute this label to a simulated vs. embedded system.

25 This is not to say that these approaches do not allow or facilitate the designer's expressing, imposing, or verifying (formal) specifications on the system/environment interaction. For example, this is the emphasis of Pnueli's advocacy of temporal logic as a tool to express temporal specifications (of a limited class) for "reactive" systems. "Reactive", in Pnueli's sense of the term, refers to programs such as computer operating systems, or airline reservation systems, which do not normally terminate, but must keep an ongoing interaction with the environment. (See, for example, A. Pnueli, "application of temporal logic to the specification and verification of reactive systems: a survey of current trends", in J.W, deBakker et al., eds., "Current trends in concurrency:overviews and tutorials", Lecture Notes in Computer Science, Vol. 224, Springer, 1986.) Note that this differs from our use of the term "reactive" in the present paper, to refer to the absence of control structures mediating the system's response to environmental stimuli. (Related to this, but distinct, is the use of the term, e.g., by Brooks, to refer to systems not making use of "internal representations" of the environment. In our opinion the notion of "representation", and consequently this use of the term "reactive", is somewhat problematic. In particular, how can one tell that these systems do not, in fact, construct internal representations, albeit in a form neither planned nor recognized by the designer ?) We note also that there is a range of other formal work besides temporal logic (e.g., in timed process calculi) intended to provide means of expressing various types of temporal specifications.

We wish to emphasize that system/environment specifications imposed by the designer are not the same thing as constraints which arise intrinsically from the embeddedness of the system. Also, there is a significant difference between imposing (and verifying) specifications on an embedded vs. formal system. This is particularly significant in the context of "safety-critical" design, where system failure can have consequences which are catastrophic not only for the system but for the people using it. For a discussion of the problematical nature of "hardware verification" see A. Cohn,"Correctness properties of the Viper block model: the second level", in G. Birtwistle and P.A. Subramanyam, eds., "Current trends in hardware verification and automated theorem proving", Springer, 1989.

particular both the behaviors themselves and a priority hierarchy are "put in by hand". As the number of behaviors increases, such ad hoc approaches become increasingly difficult to implement successfully, hence the attempt to incorporate genetic algorithms (acting on a design space) to aid in the design process.[26] However, this brings us back to our criticisms (see Section 2.12) on optimization approaches. Our use of the term "ad hoc" is not intended as a criticism of this work per se. It is intended, rather, to point out that "coherence" has not been taken seriously as a research issue in itself, requiring its own theoretical treatment.[27]

One basic problem that such a theory must deal with is how to even formulate an operational definition of "coherence". We conjecture that such a definition will be linked to the value-framework (or value system, as we shall call it in Section 5B) of the system. This will involve a kind of non-vicious circularity, since we wish the value system itself to be "coherent". (As we shall see in Chapter 7 various forms of non-vicious circularity are endemic to the topic of adaptive autonomous systems).

In some sense, "coherence" is the counterpart for embedded or embodied systems of the notion of "consistency" for a formal system. But we feel that the notion of "coherence" is much richer and more subtle. For example, there is in essence only one way to be inconsistent, namely to give rise to a contradiction, whereas there are many ways for a system to be (potentially) incoherent.[28] This is perhaps the basis for the success of formal logic. Similarly, this is perhaps at the root of the formal mathematical comparability of distinct formal systems, including model-theoretic methods. It is unlikely that correspondingly sharp modes of comparison will be available for autonomous systems. The relevance of "universality" is also in question. Thus, while there is a (unique up to equivalence) universal Turing machine which can simulate any Turing machine, it is unlikely in the extreme that there is anything like a "universal" coherent (autonomous) system. Similarly, we would be much surprised if there were to be anything in this setting corresponding to the notion of a "categorical theory"(such as the axioms for the real number system), i.e., one for which any two realizations are "equivalent" (in any of a variety of senses).

It is important to recognize that "coherence" can accommodate "inconsistency". This is a key difference between autonomous systems, which are embedded in the world, and formal systems. For a formal system an inconsistency is generally taken as catastrophic.[29] For an autonomous system, on the other hand, inconsistency need not

---

26 See, for example, J. R. Koza, "Evolution of subsumption using genetic programming", pp. 110-119, in F.J. Varela and P. Bourgine, eds., "Towards a practice of autonomous systems", MIT Press, 1992.

27 One explicit formulation of a kind of coherence condition does arise in the context of concurrent process theory, in the form of "fairness" requirements. We have also noted (cf. Section 2.6e)) the problem of maintaining coherence in the network theory setting. A complex circuit, built even according to design, has its own modes of coherence maintenance, which the designer may or may not have an understanding of.

28 The notion of incoherence is, in two ways, reminiscent of John Austin's notion of "infelicitous" utterances (see J.L. Austin, "How to do things with words", Harvard U. Press, 1975). First, the hint of embeddedness/embodiment associated to the utterance considered as a "speech act". Second, the variety of senses in which an utterance can be "infelicitous", as contrasted with simply "ill-formed".

29 However there is a tradition of work on the logic of inconsistency. (See, e.g., G. Priest, R. Routley, and J. Norman, eds., "Paraconsistent logic: essays on the inconsistent", Philosophia Verlag,

cause a problem until it is necessary for actions to be taken; even then, it is a matter of conflict resolution, possibly on a case by case basis, rather than of rooting out contradictions in the "foundations" of the system. Some examples are: inconsistent beliefs held simultaneously, inconsistent goals, conflicting values, etc. "Resolving" these conflicts may involve the making of "choices". This "decision-making" process, in order to terminate in timely fashion, may require more than "disinterested" (or "universal", or "objective") principles of "rationality"; it may require the involvement of "internalist", possibly body-based mechanisms.[30]

2.15 Non-algorithmicity

The approaches 1-4 discussed above are each algorithmic in one or both the following two senses:(i)They may view the systems that they are applied to as themselves relying primarily on algorithmic modes of operation. (ii) Alternately, they use algorithmic processes to simulate or predict the behavior of these systems, not themselves assumed to be algorithmic.

We believe strongly that biological organisms, while they may in some contexts make use of algorithms, do not rely primarily on algorithmic modes of operation (even as regards "cognitive" processes). All the preceding discussion, in particular the emphasis on emergence and the role of the value system, argues against such algorithmicity. To illustrate this in a cognitive setting: we are dubious (despite arguments to this effect in the A.I. literature) that "intuition" can be grounded in algorithmic processes. In particular, we

---

Munich, 1989.) Work in the AI community on non-monotonic logic, ATMS (assumption-based truth-maintenance systems), etc. should perhaps also be considered in this connection.

30 We are here only taking note of a vast topic, which we hope to return to elsewhere in more detail. (In the present paper, the question of "conflict resolution" figures significantly, if largely implicitly, in the design of the functional architecture schema considered in Chapter 7). We do wish to add a few remarks here, however. When we speak of "conflict resolution" within or by an autonomous system we have in mind a range of contexts, as well as time scales. For example, in the setting of biological organisms, this conflict resolution may involve natural selection acting on populations on an evolutionary time scale (see, e.g., the footnote to Section 3.1c)), or it may involve processes within an individual organism on a "somatic" time scale. In this latter context, the issue of "conflict resolution" is pertinent to "simple" organisms as well as to higher organisms such as humans. In particular, neither higher cognitive processes nor awareness, self-awareness, consciousness, etc. are necessary concomitants of "conflict resolution" as we conceive the term.

Conflicts (both conscious and unconscious) have obviously been a major focus of study in the context of human psychodynamics, and not only from a Freudian perspective. (See, for example, the work of Ainslie, referenced in one of the footnotes to Section 1.1). In speaking of the need to go beyond "objectivist" rationality, we are not referring to considerations of limited computational resources, as associated with the term "bounded rationality". Rather, we are concerned with the more basic issues of principle addressed by A. Damasio in connection with his "somatic marker" hypothesis. (See the reference listed in the footnote to Section 7.12(iv)).

In the context of moral philosophy, it has been a major theme in the work of Isaiah Berlin that ultimate values are inherently incommensurable, and are uncombinable within a single human being or single society, and that there is no overarching standard by which this conflict is rationally arbitrable. This idea has subsequently been taken up in the context of political philosophy by Joseph Raz, who refers to "constitutive incommensurabilities". (See, e.g., I. Berlin, "Two concepts of liberty", 1959; J. Raz, "The morality of freedom", 1986; as referenced in, J. Gray, review of "Isaiah Berlin: a celebration", Times Literary Supplement, July 5, 1991, p.3).

(In contrast with the "value incommensurability" discussed above, we think that "theory incommensurability" (as emphasized in T. Kuhn's work on the history and philosophy of science) is more apt to be relevant to the modeling framework mentioned in the footnote at the beginning of Section 2B than to the organization of autonomous systems themselves).

think that algorithmically-controlled search of a pre-existent "search space" (cf. Section 2.11) is fundamentally different in character from "search" guided by intuition. (To oversimplify somewhat, we think of algorithmically-controlled search as an auxiliary process to be brought into play after non-algorithmic processes have focused on "relevant" possibilities.)

The fact that a system is not itself algorithmic does not ipso facto imply that one cannot simulate or predict its behavior to "arbitrary" precision via algorithms run on a computer. The obvious case in point is planetary motion. (No one would claim that gravitational forces act via computation. Yet Newton's laws can be expressed in the form of differential equations, which can then be solved algorithmically.) However, we feel that the very nature of autonomous systems may impose limitations of principle (in contrast, say, with limitations associated to inadequate "computing power", or to sensitive dependence on initial conditions, as in "chaos theory") on the extent to which the system's behavior can be simulated or predicted. The limitations we have in mind here are not centered on the coarseness vs. fineness of "mesh" that can be used, but rather on the possibility that the whole notion of "prediction" must be recast in the present setting. Rather than being akin to the predictions that a theorist can make about the motion of a system of particles, it may be closer in character to the kinds of "predictions" made in a social setting, of one individual about the behavior of another.[31] We put this "conjecture" forward quite tentatively, but think it deserves serious investigation.

C. Neurobiological approaches to integrative organization

2.16 Approaches to be discussed

There has been a recent upsurge of interest in "integrated cognitive architectures", particularly in the fields of AI/robotics and cognitive science.[32] However, comparatively few of the recent proposals directed towards the integration of "component" subsystems have been put forward from the standpoint of cognitive neuroscience, i.e., approaches grounded in the neurobiology of the (human) nervous system. From this standpoint an abstract functional architecture is not sufficient. The "components" of the architecture must be linked to specific anatomical and physiological substrates; reciprocally, there is an emphasis on the association of functional "roles" to the various structures, patterns of connectivity, and patterns of physiological activity of the brain, both at a large-scale and at a fine-grained level.[33] In the present section we shall briefly touch upon three relatively recent neurobiological approaches which either propose or aim at comprehensive cognitive architectures. As in Section 2B, our discussion is intended primarily to help situate our own ideas, rather than to provide a review of these

---

31 This divergence as to the meaning of "prediction" may play a role in the difficulties of human/machine interface as discussed by Lucy Suchman (see L.A. Suchman,"Plans and situated actions", Cambridge U. Press, 1987).

32 For a sampling of work in the spirit of AI/robotics see, e.g., SIGART Bulletin, Vol.2, No. 4, "Special section on integrated cognitive architectures".

33 The types of correlation sought here in no way imply a naively localizationist view of brain function.

approaches. Accordingly, we will limit comments to the points of contact with our work.[34] The approaches to be discussed are:

1. The theory of neuronal group selection (G.M. Edelman)
2. Convergence zones and somatic markers (A. R. Damasio)
3. Neural symbolization (J. Kien)

2.17 The theory of neuronal group selection

In its "extended" form, this is a global theory of the brain, encompassing development, functional organization, and a brain-based theory of consciousness.[35] Our comments here will be limited to considerations of functional organization. The work of Edelman and his collaborators emphasizes an embedded, non-algorithmic view of functional organization, and stresses the need for internalist perspectives, especially in connection with the emergent processes of perceptual and conceptual categorization by the organism. The fundamental "mechanism" underlying these processes, according to this theory, is a process of selection , occurring on a "somatic" time scale (i.e., within the lifetime of the individual organism) and acting on "repertoires" of "neuronal groups". This results in the formation of "local maps", with "reentrant connections" among the local maps leading to the formation of coherent "global maps". A key role in biasing and shaping the organism's behavior, in particular its perceptual categorization, is played by a system of evolutionarily selected "value units", or "value repertoires" associated to basic survival needs.[36]

The theory of neuronal group selection emphasizes non-algorithmic, internalist perspectives, and is concerned with the avoidance of "homunculus problems". In these respects our own orientation is similar. At a more specific level, the "primary value system" which we introduce (see Sections 3.3 and 5B) bears some affinity to the above-noted value units. Also, our broad notion of "neural symbol"(see Chapter 7) can probably accommodate the above notions of "local map" and "global map", and our picture of neural symbol manipulation can perhaps accommodate the categorization or manipulation of global maps which figures in Edelman's treatment of "concept formation". Some

---

34 In particular, we will not discuss, except in passing, the specific proposals made in these approaches as to the neurobiological substrates supporting particular abstract functionalities. It should be clear from our emphases throughout the present paper on "embeddedness" (or "embodiment") and "emergent" functionality, that we do not regard such questions of substrate as purely incidental considerations. It is simply that at the present time they do not impinge directly on our own work.

It also should not be assumed that in failing to even mention other work in this very large field we are either making some sort of implicit value judgment or displaying a complete ignorance of the literature. We are exercising a deliberate myopia here, and limiting our discussion to work we perceive as having some particular affinity with our own. In light of our emphasis on non-algorithmicity, this accounts for our omission of any discussion of "computational neuroscience" or of combined AI/neuroscience perspectives, even intriguing proposals for comprehensive integrative architectures such as that put forward by Yves Burnod. (See, Y. Burnod, "An adaptive neural network: the cerebral cortex", Masson, Paris, 1989).

35 See G.M. Edelman, "Neural Darwinism", Basic Books (1987); "Topobiology", Basic Books (1988); "The remembered present", Basic Books (1989); "Bright air, brilliant fire", Basic Books (1992).

36 For an exposition of the role played by value units in perceptual categorization see, for example, G.N. Reeke, L.H. Finkel, O. Sporns, and G. M. Edelman, "Synthetic neural modeling: a multilevel approach to the analysis of brain complexity", in G.M. Edelman, W.E. Gall, and W.M. Cowan (eds.), "Signal and sense: local and global order in perceptual maps", J. Wiley & Sons, New York (1989).)

points in which our approach differs from that of the theory of neuronal group selection are: (i) Selectionist processes do not play such a dominant role. (ii) We emphasize a more elaborate "value system", including higher-order and symbolic values. (iii) The problems associated with "coherence" (cf. Section 2.14) and "conflict resolution" (including conflicts among values) play a more central role. We anticipate that "maintaining coherence", in the sense that we use the term, will require more than a single dominant mechanism such as "reentry".

2.18 Convergence zones and somatic markers

A rather different framework for a comprehensive cognitive architecture, also from an internalist perspective, has been proposed by A.R. Damasio. In this framework the "body proper" plays a central role in grounding, mediating, and guiding the "higher" cognitive processes, and even in maintaining the sense of identity of the mind/brain. Two key components of this architecture are "convergence zones" and "somatic markers".

a) Convergence zones[37]

These form the neuronal substrate of a "combinatoric" schema whereby "dispositional representations" activate and manipulate "images" ("topographic representations"). These "images", which form the contents of consciousness, are correlated with patterns of neuronal firing at distributed sites in the primary sensory and motor cortices corresponding to distinct "feature-based" sensory and motor fragments. The integration of these various fragments is achieved via the synchronous activation of the corresponding neuronal firing patterns. This activation is mediated by firing patterns ("dispositional representations") supported in "convergence zones". These are neuronal sites located downstream from the primary cortices which receive feedforward connections from the distributed primary sites and, reciprocally, form feedback connections with them. This is envisaged as an iterated, hierarchical arrangement, with several tiers of successively higher-order convergence zones. This schema serves at once as a means of organizing memory storage and as a mechanism of recall, via the synchronous reactivation of distributed feature fragments.

b) Somatic markers[38]

According to the "somatic marker" hypothesis a system of (visceral) body-based, emotionally-mediated value markers (both "positive" and "negative"), becomes associated (via the individual's nervous system) to the consequences of various courses of action, on the basis of the individual's experiences. This mechanism of somatic marking plays a fundamental role in guiding and shaping the individual's decision-making (and "intuition"). Clinical case studies by Damasio and his colleagues, in which ventromedial frontal lobe damage leads to disruption of the somatic marking mechanism, strongly suggest that in the "real world" (i.e., in the domain of "social cognition" vs. formally-structured artificial problem domains) a purely "rational" decision-making process, not

---

37 For an early reference see, for example, A. R. Damasio, "Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition", Cognition , 33 (1989), pp. 25-62.

38 See A. R. Damasio, "Descartes' error", forthcoming, 1994, Putnam, Grosset Books.

grounded by a body-based value system, often leads to infelicitous consequences. Among the mildest of these consequences is a tendency towards "infinite" dithering.[39]

The above framework places an explicit emphasis on embodiment (cf. Section 2.5c)). Concomitantly, this framework has a strongly internalist character. This is exemplified by the treatment of the organism's "rationality": as emergent, internal-value dependent, and non-algorithmic. We feel that our notion of "neural symbol" is compatible with the concept of "representation" used here, including both "dispositional" and "topographic" representations.[40] Moreover we think that there is an affinity between our picture of neural symbol manipulation and Damasio's picture of representations manipulating or activating other representations. In addition, our "higher-order value system" (see Sections 3.3 and 5B) very naturally accommodates the somatic marker system, and it plays a correspondingly important role in our approach. Again as above, one difference we see with our approach is the central role we give to the issues of "coherence" and "conflict resolution".

2.19 Neural symbolization[41]

J. Kien, in a discussion of her work with J. Altman on time-frame descriptions of behavior, makes a persuasive case for the potential relevance both to human memory and to motor organizing systems of some form of non-computerlike, embedded, context-dependent process of neural symbolization. Among the roles envisaged for such a process are: data compression (whereby original "data" can be maintained intact and reaccessed as needed, while compressed "surrogates" undergo actual "data-manipulation"); serial ordering; symbolization of relationships (and of relationships of relationships). We think that the process of "neural symbolization" which we discuss in Chapter 7 has some of the qualitative characteristics advocated by Kien.

We agree with certain broad features of the multiple time-frame model, in particular the emphasis placed on the issue of behavioral choice, in connection both with memory activation and with motor planning. However, once again we would like to give greater emphasis to questions of "coherence". Such questions could, for example, be posed in conjunction with the "mechanism" for behavioral choice proposed by Kien and Altman. This is a highly distributed form of "decision making", involving "consensus" among strongly interconnected "processing networks" of neurons (modeled as attractor

---

39 Consistent modes of deviation from normative rationality by normal individuals (without brain damage) have been noted in the cognitive science literature, but have been interpreted in a pejorative fashion. (See, for example, A. Tversky and D. Kahneman, "Rational choice and the framing of decisions," 1986, reprinted in G. Shafer and J. Pearl, eds., "Readings in uncertain reasoning", Morgan Kaufmann, 1990.

40 Our "neural symbols" are not abstract entities, with form divorced from content, subject to purely formal syntactic manipulation, as in the case of the symbols in a formal language. In particular, we think that our notion of "neural symbol" in our sense can accommodate both the "rich images" emphasized by Damasio and "schematic images" emphasized by Johnson and Lakoff (cf. the references in the footnote to Section 7.12(iv)).

41 See J. Kien, "Remembering and planning: a neuronal network model for the selection of behavior and its development for use in human language", pp. 229-258 in K. Haefner, ed., "Evolution of information processing systems", Springer, 1992.

neural networks)[42] ; in this picture making a decision corresponds to the process of reaching a stable state (not necessarily a fixed point) of the associated dynamical system. From the standpoint of "coherence" the relevant questions would include: (i) What spatio-temporal patterns of stable states correspond to "coherent" behavioral choices ? (ii) In this particular dynamical system setting, what mechanisms (not requiring the intervention of the modeller during "somatic" time) insure that coherence is maintained ?

42 In the sense of D. J. Amit. (See, for example, "Modelling brain function", Cambridge U. Press. 1989.

3. Characteristics of adaptive autonomous systems: what must a mathematical framework accommodate ?

In this chapter, we will seek to isolate some of the characteristic features of systems that one would regard as adaptive and autonomous. We want to explore the following types of question: (i) What kind of "ontology" must a mathematical framework for such systems accommodate? (ii)What kinds of problems of integrative organization will mathematical approaches need to address ? (iii) What kinds of mathematical approaches are possible, and what are their limitations? In the discussion below, the notions introduced, such as "goals", "values", etc. will be used with their intuitive meaning, rather than in any special technical sense. We will examine these various notions in more detail in subsequent chapters.

A. Characteristics of adaptive autonomous systems

3.1 The basic requirement : relative adaptation to the external environment

a) The environment

By definition, any interactive system, even a Turing machine, has an environment of one sort or another, even if this environment is extremely schematic (a tape, in the case of the Turing machine).[1] Depending on the nature of the system, its modes of interaction with the environment may take various forms, for which a variety of descriptive "metaphors" may be to a greater or lesser extent apropos, often concurrently. In some cases it may be appropriate to speak of "inputs" and "outputs", which may be (sometimes simultaneously) "signals", "substances", "actions", etc. (reading or writing symbols on the tape, in the case of the Turing machine).The environment, as well as the system, may be "modified" in its "state" or in its activity as a result of this two-way traffic; in particular, intermediate outputs may influence subsequent inputs (as in the case of "reactive" systems in the sense of Pnueli, mentioned in Section 2.14, footnote). In some cases it may be appropriate to think of specific, discrete "channels" of communication, either fixed or time-varying. In other instances, as in the case of an organism, or even of a cell embedded in its milieu, where the interface between system and environment is a complex membrane of selective and modifiable permeability, such a picture would be highly caricatural. Moreover, even in the context of discrete channels, "communication" need not be point-to-point targeted, but may be of a diffuse "broadcast" nature, and the "tokens" of communication may be heterogeneous in structure and polymorphic in their effects. Also the effect of "signal reception" may involve a complex triggering and transduction process rather than a simple passage from outside to inside. In addition, the system may be able to exercise considerable selectivity in its receptiveness to "inputs".

---

1 An autonomous system may to some extent decide what aspects of its embedding milieu constitute the "environment". (See the footnote to Section 2.4).

b) Maintaining specified relations with the environment

Many artificial systems are supposed to maintain a certain number of relations with the environment. For example, a reservation system should be able to deliver tickets and schedules to the environment, which, in this case, is a file of customers; a plant transforms certain industrial inputs into outputs which should be adjusted to the customer's demand; a transportation system should efficiently transport people... Abstractly, these systems are "reactive" (in the sense of Pnueli, noted above); that is, they receive inputs from the outside and produce outputs modifying the environment, but they are intended to produce outputs in such a way as to maintain given relations which are specified by the designer of the system. These specified relations can be very precise (e.g., in an air traffic control system, two airplanes should not collide), and may be expressible via mathematical equations or inequalities, or via some formal specification language.(However, it should be noted that expressing the specifications, enforcing them, and verifying or monitoring that they are being or will be met are distinct matters). On the other hand, they can be much less sharp, possibly admitting of varying degrees of compliance. Towards one end of this "sharpness" spectrum are relations which can be specified only using natural rather than formal language, and where compliance or noncompliance is subject to interpretation. Many examples of such relations arise in "normative" systems, such as legal systems or systems of social conventions. In such normative systems even the question of whether a particular situation falls within the scope of a given "law" or "rule" is subject to conflicting interpretation. We shall return to this question of sharpness in a later chapter, in connection with "vague" specifications.[2]

c) Natural systems:maintaining a state of relative adaptation with the environment

A natural system, like a cell or an animal, which is not subject to designer-imposed specifications, relates to the environment in a rather different manner. Being embedded in its environment it must, to preserve its very existence and its autonomy, maintain a state of relative adaptation. The term "relative adaptation" (as used by K.F. Liem[3] in the context of form-function evolution) is intended to convey several important connotations: (i) In a rapidly changing and rich environment, it may be inadvisable, if not impossible for the system (on a somatic time scale; or for a population, on an evolutionary time scale) to implement a "strategy" of maintaining fixed relations with the environment. Instead, the system will need to adapt to the environment, which means, in particular, that it may need to vary the relations it maintains with the environment (cf. the reference to "dynamic equilibrium" in Section 2.6d)). (ii) This adaptation is "relative" in two senses. First, even supposing there were some intrinsic utility function correlated

---

2 It makes sense to request an independent review of a given judicial opinion, but surely it is preposterous to ask for a "proof of correctness". Speaking rather informally here, our view is that the degree of "vagueness" of specificity (not intended here as a pejorative!) associated with biological systems lies somewhere towards the middle of the spectrum between sharp specificity and the looser specificity associated with normative systems.

3 Liem, K.F., Adaptive significance of intra- and interspecific differences in the feeding repertoires of Cichlid fishes , Amer. Zool. 20 (1980), pp.295-314 . Liem presents a picture of the organism as a "network of interacting constraints", or a "form-function complex". Interpreting these as design solutions, he emphasizes the integrative character of the "designs", so that modifications with respect to some particular design feature may enhance some capabilities of the system while at the same time diminishing others. Apropos "structural integration" see also S.J. Gould, "The evolutionary biology of constraint", Daedalus, Spring 1980, pp. 39-52.

with "degree of adaptation" of behavior, the environment might be changing too rapidly for the system to ever succeed in optimizing this utility. More importantly, there is no reason to think such a utility function exists. The environment, rather than setting fixed "problems", affords various possibilities to the system, and behaviors which take advantage of some of these preclude taking advantage of others. Moreover, what is needed is not individual behaviors but an integrated repertoire of behaviors. Thus, depending on the degree of stereotypy of its environment, the system, in order to maintain a degree of relative adaptation adequate for survival and autonomy, may require a capacity for learning, or even for context and history-dependent behavioral-choice. (A potentially helpful analogy: rather than determining how to play a particular game, the question is which game to play).

3.2 Basic constraints [4]

a) Constraints of time and metabolism.

The functioning of a system, whether it acts purely reactively or draws upon higher-order control, requires time and resources. In most modeling, the resources are forgotten (a formal automaton does not need food or fuel !). For embedded systems, whether machines or animals,... the metabolism (or energy) constraint and the time constraints are of first importance. Indeed, at least for "lower" animals, much of their time is devoted to searching for food, which is necessary to maintain their metabolism within appropriate limits.

Moreover resources are scarce. This means that various subsystems of the system will need to "compete" for access to resources, either external or internal, including metabolic resources. Certain of these resources may be needed in order to manufacture other resources not directly available from the environment. The system, if an organism, may also need to compete with other organisms for external resources. Nor is it simply a matter of "competition". Scarcity of resources may compel coordinated activity, or "cooperation" by subsystems.[5]

---

4 As we shall elaborate in Chapter 4, the primary distinction we wish to emphasize is between the notion of "specification" and the notion of "constraint". However, one could also note distinctions among constraint "types". For example, certain constraints are physical laws, which any embedded system is subject to simply by virtue of being physically embedded. Others are of a more conditional nature. For example, the "metabolic" constraint introduced in this section is not a physical law, but must nevertheless be satisfied if the system is to remain alive. The same holds for other "homeostatic" or "safety" constraints.

5 J. E. Mittenthal and his collaborators, in their work on principles of integrative organization in organisms, have stressed the driving role of resource limitations in shaping the specific organizational structures that emerge. (See, e.g., J.E. Mittenthal, A.B. Baskin, and R. E. Reinke, "Patterns of structure and their evolution in the organization of organisms: modules, matching, and compaction, pp. 321-332 in J. E. Mittenthal and A.B. Baskin, eds., "Principles of organization in organisms", Proceedings Vol. XIII, Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, 1990.) A nice discussion of this, in the context of the integrative organization of *E. coli* is given in J.E. Mittenthal, M. Clarke, and M. Levinthal, "Designing bacteria".

We use the terms "competition" and "collaboration" metaphorically, and do not intend to suggest "rational agents".

b) "Spatial" constraints.

Additional basic constraints on embedded systems are associated with space. Some examples: First, the system (say, an organism) may move in its environment. Second, the various subsystems of the system must act in spatial coordination with one other, taking into account their positions, shapes, contiguity relations, etc.

3.3 Goals, values, and optimality

a) Goals

An autonomous system has its *own* goals. By this we mean that its goals are not fixed by a designer, but are emergent properties of the system in its interaction with its changing environment.[6] By "own goals" we also mean that the complex forms of behaviors are not a priori pre-specified and "pre-wired", but that they are emergent. In particular, there is not a fixed repertoire of behaviors, organized into a pre-set priority hierarchy. The term "goal" as used here is very broad. For example, it need not be tied to any "awareness" or "consciousness" or "volition" on the part of the system. Nor need it convey any intentionality insofar as this is directed towards any specific object or aim or end external to the system. The minimal goal (or proto-goal) is implicit and system-referred[7]: to maintain itself as a system. This requires maintaining a state of relative adaptation with the environment and, in particular, maintaining metabolism within safe bounds. Outward-directed goals emerge as conditional concomitants of the system's modes of fulfilling (and of giving explicit form to) its implicit proto-goal, and these modes are themselves conditioned upon the system's internal organization and "resources", and on the state of the environment.

"Emergent" does not mean emergent out of nothing; there must be a kernel, or ground, starting from which the system, in interaction with its environment, is able to "bootstrap" itself. This kernel includes a physically embedded structural/functional substrate which supports mechanisms to induce and sustain the fulfillment of the proto-goal. We shall, somewhat metaphorically, refer to this structural/functional substrate as the "basic circuitry" of the system, and shall make suggestions below as to what types of "components" it may minimally need to incorporate. We shall also propose a possible realization of the "inducing and sustaining" mechanisms: namely, a "primary value system" embodied in the "basic circuitry".

---

6 This characteristic of "autonomy", suggests that our use of the term is different from that of designers who seek to build "autonomous robots". That is, is it possible to arrange the design so that the system is autonomous in our sense, yet at the same time meets specific goals of the designer (e.g.., specific tasks to be performed, or ends to be achieved in the environment)? Perhaps this is to some extent realizable in the form of additional proto-goals built into the "circuitry" and "value system". However, it is unlikely that a specific task, such as driving a nail into a wall at a specified location on the wall can be "emergent". More likely what is desired is a "quasi-autonomous" system with the following characteristics: (i) It can act autonomously (i.e., "take care of itself" and "not cause damage") when not required to perform specific tasks. (ii) When required, it accepts and is able to interpret and execute (in a context-dependent fashion) instructions (expressed in a high-level language) specifying the task to be performed but not the details of execution.

7 Making a rather strained analogy with a formal system setting, the idea of system-referral vs. external-referral, is akin to a recursively- defined condition on a function or process.

A couple of points of clarification: (i) The "basic circuitry" (certainly the particular form/function complex comprising it) is not "canonical", and may vary from system to system.[8] Moreover the basic circuitry in any individual system need not comprise the entire "circuitry" of the system, and must be integrated with the rest of the system organization. In particular, the "basic circuitry" should not be confused with the "initial circuitry" of the system, the latter being the initial state (at "birth" or time of "construction") of the overall "circuitry" of the system. (ii) The circuitry and value system are not static. Over time (i.e., the "somatic time" of the system), both the "circuitry" and the value system, in interaction, will undergo modification, and may or may not reach a quasi-steady state, or pass through a succession of such states. In particular, the value system, which initially consists only of the primary value system, will expand to incorporate "higher-order" values, as well as "symbolic" values. These modifications of the "circuitry" and of the "value system", which serve to enhance the adaptiveness of the system, involve processes of learning and memory. To an extent, the substrates supporting these processes may themselves be emergent, and undergo modification over somatic time. Referring to the "time scale" terminology discussed below, we shall (essentially as a tautology) assume that the "initial circuitry" and "initial value system" of the given system are the result of evolution (resp., the designer) working over evolutionary time (resp., design time). The other constituents of the value system will be emergent. Here, and elsewhere, when we speak of "emergence" we shall mean emergent in somatic vs. design time, unless we explicitly state otherwise.

b) Time scales

We like the idea emphasized by Edelman (see Section 2C) of introducing distinct time scales. In the case of a biological organism these include: evolutionary time, developmental time (i.e., the time of passage from zygote to adult, or to some intermediate state of maturity), and somatic time (the lifetime of the mature individual, or relevant epochs of varying scales within the individual's lifetime).

Inasmuch as our focus is on autonomy and adaptiveness, and not on "life", we do not wish to emphasize reproduction (or self-reproduction) as an issue, and certainly not in the context of artificial systems. Also, we wish to allow significant structural changes during the entire lifetime of the system. Thus, we shall not be concerned to distinguish developmental time, and shall generally subsume it under somatic time. Moreover, in the context of artificial systems, we shall substitute "design time" for "evolutionary time". This is the time during which the designer may generate and test a (parametrized) family or "population" of designs. We picture this design time not so much as a distinct time *scale* , but as a distinct time axis, interleaved with the somatic time axis. During design time any testing, modification, or selection is done by the designer on a population of systems or of system designs. During somatic time the designer does not intervene, and there is only the individual system and not a population. The only "exploration" is that done by the system, and the domain explored is the environment, via direct interaction, rather than a "design space". Such interaction may result in (beneficial or harmful)

---

8 In particular, we do not mean to suggest anything like a canonical functional microcircuit serving as a repetitive building block for the system, as has been proposed for neocortex (see, e.g., R. J. Douglas, K.A.C. Martin, and D. Whitteridge, "A canonical microcircuit for neocortex", Neural Computation 1, pp. 480-488, 1988).

modification of the system, including modification of the "circuitry" and of the "value system". Such modification may, in particular, take the form of "learning". If the system has the capacity for "planning" this provides, in somatic time, a kind of "virtual" population, as a sort of partial "surrogate" for the actual population of design time.

c) The basic circuitry

We imagine that the "basic circuitry" would, at a minimum contain mechanisms which manifest sensitivity to alterations in the internal state of the system, as well as mechanisms capable of restoring the internal state to a "baseline", or "dynamic equilibrium" level. This may require more than a simple feedback loop; e.g., it may require significant internal structural or functional reorganization of the system achieved via internal effectors acting "mechanically", in conjunction with cascades of "chemical" processes, Moreover, the "mechanical" processes may themselves need the support of the "chemical" processes to fuel their activity. In addition, the activity of one set of effectors or processes may give rise to the need for additional mechanical and/or chemical activity, for example, to eliminate waste products. Part of the "job" of the effectors may also be to alter the situation of the system in the external environment, in particular via locomotion. The system will also require "interfaces" with the external environment, in order to take in resources and to eliminate waste products. We do not expect the interface to be a purely passive conduit. Hence there will be the need for effectors to physically control the activity of the interface, as well as mechanisms sensitive to alterations in the state of the interface. The state of the interface should undergo some (localized or diffuse) alterations induced by physical stimuli impinging on it from the external environment, so that the system has the possibility of taking account of "resources" or "risks" in the environment. (We expect the basic circuitry to act predominantly in an "autonomic" fashion; i.e., independently of direct control intervention by other parts of the circuitry, as in the case of the autonomic nervous system. By this we mean that the remainder of the system does not directly monitor, and hence does not directly regulate the state of the basic circuitry; rather, it influences the behavior of the basic circuitry only by influencing the state of the external environment of the basic circuitry).

We have spoken of "sensitivity" rather than of "sensors" so as not to appear to suggest some apriori modes of categorization by the system of the sensory "inputs". We expect this categorization to be an emergent process. Similarly, the sensitivity may be of a diffuse vs. sharply localized or sharply specialized character. The system need not be sensitive to precisely where a signal has impinged, nor need the system cleanly segregate signals of different physico-chemical type. Nor need the signals result in elaborate internal representations of the environment; for example, photosensitivity need not imply a sense of vision. The sensory and effector apparatus may become refined and elaborated over time (evolutionary/developmental/somatic).[9] This refinement will typically go along

---

9 We do not rule out multifunctionality. Thus, the same physical component or activity pattern may serve both sensory and effector functions.

In an interesting dissertation, Peter Cariani seeks to develop a framework in which to distinguish between computation, measurement, control, and non-symbolic functionalities. He proposes a taxonomy distinguishing between "nonadaptive" devices ("formal-computational" and "formal-robotic"); "computationally adaptive" devices (e.g., neural nets or genetic algorithms); and "semantically adaptive" devices ("evolutionary" devices). The devices in this last class adaptively construct their own sensors and

with an elaboration and reorganization of the overall structural/functional architecture of the system. The sensory apparatus need not function in a purely "passive" manner. Rather, the system may actively "explore" for "relevant" signals in the external or internal environment.

It is conceivable that, from the standpoint of a strictly formal/logical vs. embedded system, a complete list of the logical "components" required for (one of the variants of) a "basic circuitry" may be very brief (as is the case, say, with von Neumann's "self-reproducing automata";see Section 3.7). However, when one considers an actual embedded circuitry, the situation may be quite different. As hinted at in the above partial enumeration, the inclusion of any one component gives rise to requirements for additional supporting substrates or processes, which in turn generate support requirements of their own, etc. By the time one attains "closure", or "self-containment" (in dynamic equilibrium with environmental resources), one may have a very complex system. For example, a cell, even if we disregard apparatus specifically associated with replication (the cell cycle) is tremendously complex. Nor would we say that a virus (which draws on a host cell's reproductive apparatus) is "simple".

d) Values

An adaptive system, embedded in a complex environment, cannot try out actions completely "at random" all the time. (Here "at random" may, but need not, connote "stochasticity". The emphasis, rather, is on "non-directedness"). Even if certain actions are done at random, as may particularly be expected early on in the "lifetime" of the system, there should be a system of values or rewards which has the effect of retaining "valuable actions" for the system. We envision this value system as an outcome-based biasing system, rather than as an encoding or representation (explicit or implicit) of pre-set behaviors; indeed, in general, the value system will have no way of "knowing" (nor need there in principle be a way of knowing) what the (a posteriori) beneficial behaviors or goals actually are. In particular, the value system should be consonant with situatedness, or context-dependence, of action, and with behavioral choice. As discussed above, initially the value system consists exclusively of the primary value system, which is associated with the minimal, proto-goal, and is realized via the basic circuitry. Over the course of (somatic) time, the value system can expand, with certain new values appearing, and others disappearing or being modified, depending also on changes in the environment. As emphasized above, these values are emergent, and not imposed by an external agent.

In speaking here of a "biasing system", we shall not (for purposes of the present chapter) seek to distinguish between: the system for generating the "biases", the physical substrate which supports the "biases" thus generated, or the "biases" themselves. (For that

---

effectors contingent on their performance. Cariani links these three classes of devices to different forms of Robert Rosen's notion of "emergence relative to a model", i.e., the deviation of the behavior of a natural system from a model of it. (This is to be contrasted with "computational emergence" and "thermodynamic emergence").According to Cariani, the nonadaptive devices do not yield emergent behavior; the computationally adaptive devices require only adjustment of model parameters (or state transition rules) to recover predictability; and semantically adaptive devices require the continual introduction of new observables in order to regain predictability. (See P. Cariani, "On the design of devices with emergent semantic functions", Ph.D. Dissertatation in Advanced Technology, State University of New York at Binghamton (1989); R. Rosen, "Anticipatory Systems", Pergamon Press, 1985.

matter these distinctions, certainly that between a bias and its substrate, are quite murky). In particular, the primary value system involves all the above, including certain "pre-set" biases determined in evolutionary time (or in "design" time). Nor do we mean to suggest a uniformity or homogeneity of "mechanism". As one example, even in the case of the primary value system, certain of the primary values, such as those directly linked to homeostatic constraints, may involve error-correcting mechanisms of the traditional "feedback" variety. Others, probably constituting the general case, will have more of the "unspecific" or "diffuse" character emphasized in the preceding paragraph.[10] We should also note that when we speak of individual "biases" or "values" as entities in their own right, we do not intend this in any literal sense: first, because a particular "value" makes sense only as part of an overall system; secondly because, as emphasized above, values in the sense we have in mind tend to manifest a kind of generalized, or diffuse vs. hard-edged quality.

e) Non-optimality

Reiterating the point made in Section 3.1c), the "optimality metaphor" does not apply to the workings of an adaptive system. In general an adaptive system does not try, either through its own "intent" or through the nature of its underlying "design", to optimize anything in particular. There may be many conflicting, incommensurable criteria of utility present concurrently, rather than a single overarching criterion.[11] When the system , or some subsystem, does optimize a particular "utility function", this utility criterion may readily vary from one context to another. Moreover, even when an behavior does in fact "optimize" some criterion, we expect this criterion itself to be emergent from the interaction of the system with its environment and not, in general, prefigured in the "design" of system.

3.4 The need for control: achieving and maintaining coherence

All adaptive systems (whether natural or artificial) will tend to be rather complex. Even to maintain themselves as systems, let alone to accomplish a specific set of outward-directed goals or tasks, they will need to integrate the activity of multiple interacting subsystems. Moreover, this internal organization of structural resources into functional subsystems may be dynamic rather than static. In addition, distinct "levels" of subsystem "decomposition" may be superimposed. (This is illustrated most readily in the case of societal organization: among the superimposed levels are the political, legal, economic, familial, etc.)

---

10 In this respect they will be more akin to the "value units" or "value repertoires" appearing in the work of Edelman and his collaborators (see Section 2C). In this setting there is a corresponding "diffuseness" vs. sharp localization in the patterns of connectivity of the value repertoires with the other parts of the nervous system. Diffuseness of effect also figures in conjunction with the quite different, emotion-based somatic-marker schema of Damasio (see Section 2C). Here the diffuseness is associated with the central autonomic effectors, notably the amygdala, which can activate somatic responses in viscera, vascular bed, endocrine system, and nonspecific neurotransmitter systems.

11 We think that the "optimality metaphor" is fundamentally out of place in this setting. In particular we don't think that any aid or comfort is to be drawn from the techniques of so-called "multi-criteria optimization", e.g., forming new utility functions from weighted averages of component utility functions, or drawing on Pareto optimization ideas from microeconomic theory. We note in passing that in Pareto optimization theory one does not have a canonical way to choose from among the many Pareto optima.

However, as we have emphasized in Sections 2.6d) and 2.14, a collection of subsystems put together willy-nilly cannot be expected to yield "coherent" behavior, let alone behavior continually maintaining a state of adaptation with the environment. Even if the subsystems are themselves individually "coherent", they may very well tend to compete in an anarchical way for access to various resources. (This need not be limited to a societal organization; to take a very simple biological example, antagonistic muscles could attempt to work simultaneously).

One is thus faced with the question of "control": namely, what is the means by which the system achieves and maintains coherence in a context of emergent functionality? We feel that this conception of the problem of control will require that one go well beyond the traditional perspectives on control engineering discussed in Section 2.6d). Even in the setting of artificial systems, the various subsystems may each have *several* naturally associated "utility functions". Correspondingly, the design of the control system may involve complex tradeoffs.[12] In particular, it will be necessary to reexamine the extent to which control can maintain its traditional "supervisory" character. In a setting of emergence and situatedness, a putative supervisor (associated with any aspect of the overall organization) might, in principle, have no way of "knowing" (certainly not to a high degree of "specificity") what task, state-of-affairs, internal state, etc., should definitely be achieved or definitely be avoided by that portion of the system under its influence, let alone what behavior should be "induced" so as to bring this about. We think that a more radical revision of viewpoint will be needed here than simply to jettison "instructionist" approaches in favor of "selectionist" ones.[13] We anticipate that a richer space of possibilities will need to be considered than the points on the one-dimensional instructionist-selectionist axis.

3.5 What else is needed?

We have discussed above in intuitive language, but still at a rather abstract level, some of the basic components that may be expected to figure in the organization of adaptive autonomous systems. Have we omitted anything essential ? We shall give three different answers from three different perspectives: (i) Undoubtedly yes. It is unlikely that we have not omitted some basic component. (ii) The question makes no sense. There is no such thing as a minimum set of "logical" components that are common to all adaptive autonomous systems, and certainly no such thing as a set of necessary and sufficient components. In particular, since we are dealing with systems with emergent functionality, it makes no sense to discuss the "logical" or formal level in isolation as a separate level. Everything is dependent on the actual concrete embedded system (see Section 2.8). (iii) Perhaps not. All specific examples that come to mind of adaptive

---

12 A separate question which we shall not take up here, and which is perhaps relevant primarily in the design (i.e., artificial system) setting, is that of incrementality of control design. For example, one may not want, when "gluing together" subsystems , to have to do substantial internal "rewiring" of the subsystems. In that case, one must constrain the control design so as not to permit (or require) it to make use of detailed knowledge of what is happening internally within the subsystems.

13 A couple of difficulties facing a "pure" selectionist approach are: (i) Certain contexts may simply not afford either the actual or virtual populations on which selective processes can act. (ii) Time constraints may not afford selective processes adequate time to act.

autonomous systems appear to present these "basic" features in one form or another. Admittedly, they individually present additional features as well, which may depend markedly on the particular type of system considered. But perhaps these features are secondary concomitants (and in some cases even "logical consequences" of) the "basic" features, serving simply as tools to implement or enhance the "basic" features. As examples of these "secondary" features , we can cite "associative learning", "categorization", "planning", "abstraction", "symbolization", ...

We do not feel impelled to make a choice from among these perspectives, of which (ii) and (iii) are surely the more interesting. We shall however, have a bit more to say along the lines of (ii) in Section 3.7, and shall discuss some of these "secondary" features in Chapter 7, where we explore some elements of an integrative functional architecture schema.

B. What can one expect from a mathematical theory of adaptive autonomous systems ?

3.6 Is such a theory possible?

The main question is "Can one hope to develop a mathematical theory of adaptive autonomous systems, and if so, at what levels of description ?". This question is not easy to answer - as probably it was not easy to answer what kind of mathematics, if any, could be used to describe a falling body or the planetary motions in the fourteen century, or to describe the action of heat and atomic motions in the middle of in nineteenth century. As a step towards addressing the question, we begin by taking note of some potential obstacles to the development of a mathematical theory. With some degree of arbitrariness, we divide these into difficulties of "practice" and difficulties of "principle".

a) Difficulties of practice

(i) Complexity

One possible objection to the relevance of mathematics could come from the "numerical"complexity of the organisms to be described. For example, a human brain has on the order of 10^{11} neurons making, on average, several hundred (and some on the order of 10^4) synapses. But a priori, this is not a compelling reason. Statistical mechanics can describe systems of 10^{23} molecules in a rather satisfactory way. To this the skeptic may reply that this is not the point, since the collective properties that emerge on a statistical mechanical scale actually simplify the problem. The issue is not numerical complexity but the complexity of the middle ground, "organized complexity".[14] Here we readily admit the difficulty of the problem, but not it's impossibility. New mathematics may well be needed, but this was never in doubt.

---

14 In the sense, say, of Weaver (cf. W. Weaver, "Science and complexity", Am. Scientist 36, pp. 536-544, 1968). For systems lying in this middle ground (i.e., systems which are "large", but not large enough) the simplifications associated with passing to infinite-limit approximations (e.g., via statistical mechanics, diffusion approximations, the central limit theorem, etc.) may not be available.

(ii) Degree of detail

A related difficulty centers on how much of the actual fine-structure of the system a mathematical treatment can hope to incorporate. To take an example, we agree that it is out of the question to describe in detail all the ongoing biochemical activity in even a single cell or, for that matter, the detailed biophysical activity at a single synapse. But to insist on such detail is to argue for reductionism not only in principle but in practice. In fact, for many purposes, the underlying organization may best be revealed by passing to a more macroscopic level, where the "relevant" variables may be few in number, such as thermodynamic variables, or large-scale dynamic modes.

(iii) Empirical substrate

Whether one regards the primary goal to be the understanding of actual systems, or whether one views the development of a theory as an end in itself, it will surely be necessary to turn to "model" systems for guidance. For the foreseeable future this means biological systems. But the data here are at the same time too much and too little. On the one hand, the available information is too extensive and rich to even sift and integrate properly, nor is it a priori obvious which of the manifold "phenomena" should actually figure in, let alone be an "explanatory focus" of a theory. On the other hand, no matter what system one selects there are fundamental gaps in what is known, and one may question whether there is an adequate empirical substrate to support the development of a theory.

We agree that this presents serious difficulties for *modeling* . Indeed, it is the source of the jarring mismatch in level of sophistication that one on occasion encounters within mathematical models proposed in the neurobiological literature. We are referring here to the introduction of detailed and mathematically sophisticated treatments in a setting where the qualitative mechanisms are still poorly understood, and the main concepts have not been precisely analyzed.

However, this need not be an obstacle to theory development, even if one of the goals of the theory is to assist in the construction of mathematical models of particular systems of interest. Moreover, with historic hindsight, the present situation is not qualitatively distinct from that facing would-be theorists of past epochs.[15]

b) Difficulties of principle

(i) Contingency

This is at the core of the ambivalence felt by many biologists towards the role of theory in their subject.[16] On the one hand, in part as a reaction against vitalism, and in part because of the great successes of molecular biology, the field has been fervent in its embrace of the Cartesian machine metaphor, with the result that the field is far more

---

15 Enrico Bellone (see "A world on paper: studies on the second scientific revolution", MIT Press, 1980), depicts the multiplicity of conflicting and shifitng conceptions by working scientists of the period 1750-1900 as to what constituted a properly "Newtonian" or "mechanistic" approach, as well as what was the proper relation of the "paper world" of mathematical theory to the world of experience. The difficulties and dangers in attempting to construct major correlations among seas of widely disparate data are well brought out in Bellone's essay (in the above collection) on the researches into electricity and magnetism circa 1750 by Jean Antoine Nollet (see "The empirical ocean and the conjectures of abbe Nollet").

16 This ambivalence is well described in R. Rosen, "Life itself", Columbia U. Press, 1991.

mechanistic in outlook than, say, the field of physics. Here the disdain towards mathematical theory is tied to objections of the type discussed above, such as "complexity", or to the argument that the absence of mathematical theory has not as yet led to any slowdown of progress.

On the other hand, there is a growing receptiveness towards mathematical theory on the part of biologists concerned with integrative organization, who would at the least be pleased to have available formal languages appropriate for describing such integrative organization.[17] However, even here there is a sense of misgiving, to the effect that the subject matter is inherently unamenable to theory. As a counterpoint to the mechanistic determinism emphasized above, there is at the same time a sense of inherent "unlawfulness", associated to the contingent, "unentailed" character of evolution via natural selection.

This is a concern which we do not dismiss. However, rather than regard such "contingency" as a barrier to any possibility of theory, we view it as a signpost, pointing towards ways in which our traditional conceptions of what constitutes a "theory" will need to be widened in order to accommodate the subject.[18]

(ii) Non-algorithmicity

A related concern is the inherent non-algorithmicity of adaptive autonomous systems, that we have ourselves emphasized in both this chapter and the preceding one, and which goes counter to the prevailing emphasis of "strong AI".[19] After all, is not mathematics in some sense the realm of algorithmicity ? Also, won't non-algorithmicity wreak havoc with any attempts at a *predictive* theory ?

We shall return to this issue of non-algorithmicity at greater length later in this paper, but will make a few remarks here. To begin with, most of the mathematics of the 19th and 20th centuries has not been algorithmic in character or in focus (leaving aside altogether the patently non-algorithmic character of the processes of mathematical discovery and invention). The current resurgence of emphasis on algorithmicity has more to do with the prevalence of the computer and computation in our society. As regards "unpredictability", algorithmic processes are themselves at risk: the study of dynamical systems (where the dynamics is fixed by rigid rules) leads to chaotic behavior; the

---

17 Such receptivity may be found, for example, in cognitive neuroscience, and in systems neuroscience more generally. However it not limited to these quarters. For example, S. Brenner has emphasized the need of biology for a theory of "elaborate systems" (cf. Judson, "The eighth day of creation"). See also the work described in the reference listed in the footnote to Section 3.2a).

18 An interesting perspective on the tensions and interplay between "lawfulness" and "contingency" is the settings of evolutionary and developmental biology may be found in the work of Stuart Kauffman (see S.A. Kauffman, "The origins of order: self-organization and selection in evolution", Oxford U. Press, 1993).

In a related vein, we feel that in the contexts we are concerned with it is inappropriate to regard "determinism" and "stochasticity" as the only possibilities, inasmuch as this leaves no avenue for "choice" to enter except as a form of randomness. This strikes us as highly artificial. At the same time, we are not convinced that the "solution" needs to be tied to some form of "wave function collapse", as associated with quantum measurement theory.

19 To clarify, we do not claim that algorithmic processes cannot play a significant role, in particular in connection with higher cognitive processes. However, even here we think that there is an intertwining of algorithmic with non-algorithmic processes, with the latter having "primacy". At a minimum, theremust be a non-algorithmic substrate to support algorithmic processes. Moreover, we expect that the process of choosing from among several algorithms is itself non-algorithmic.

definition of Turing machines leads to undecidability and unpredictability (although the functioning of a Turing machine is perfectly deterministic and rigid). Moreover, as we have suggested in Section 2.15, the relevant issue in regard to adaptive autonomous systems is not that of "predictability" vs. "non-predictability"; rather, it is the elucidation of the character of the predictions that can in principle be made, either by an external observer, or by one system of another (or of itself). To the extent that this lies beyond the scope of traditional conceptions of "predictability", we expect this will be of a much subtler character than the unpredictability attendant upon algorithmic systems. As with the issue of "contingency", we regard this as an opportunity to widen our conceptions rather than as a deterrent to inquiry.[20]

3.7 Historical background

To help provide perspective on the role of mathematical theory, we shall briefly discuss three comparatively recent (beginning circa 1930) efforts at developing serious mathematical approaches to subjects outside the traditional "exact sciences".

a) Turing machines and computer science

According to the Church - Turing thesis, a Turing machine is the mathematical formalization of the heuristic notion of computation or algorithm, or of systematic, purely mechanical manipulation of symbols. So, in some sense, a universal Turing machine is a caricature of the most general computer, an extremely simple caricature where the only operations allowed are copying and erasing. Though, as everyone is aware, beyond a very abstract functional level, Turing machines bears little resemblance to actual computers, and the formal languages that they "make use of" are extremely far removed from any concrete computer language.[21]

The remarkable confluence of work in the 1930's leading to this formalization in fact stemmed from concerns in mathematical logic and foundations of mathematics (e.g., associated with Hilbert's program of "meta-mathematics"), rather than computer science which, it is fair to say, did not at the time exist as a discipline. That it, in fact, turned out to be possible to give a mathematically precise characterization, *not depending on the particular formalism used*, of the class of processes that can be carried out purely by mathematical means was considered miraculous by Gödel.[22]

---

20 We shall also need to look for other relvant notions of compexity than those associated with an inherently algorithmic framework,such as "computational complexity" or "Kolmogorov complexity".

21 In particular the "state" space of a Turing machine need not bear any canonical relation to the "space" of states of an actual computer. It is not an "absolute" invariant of the the computational process, but a model-dependent one; i.e., it depends on the use of a Turing machine model.

22 The equivalence is between the class of general recursive functions (in the sense of Gödel-Herbrand), the λ-definable functions (in the sense of Church's λ–calculus), and the functions computable via Turing machines. Thus, Church's thesis was initially stated in terms of λ-definability.

Gödel's view was publically expressed in Remarks before the Princeton Bicentenial Conference on Problems in Mathematics-1946- (see M. Davis, ed., "The Undecidable", Raven Press, 1965):

> "Tarski has stressed in his lecture (and I think justly) the great importance of the concept of general recursiveness (or Turing's computability). It seems to me that this importance is largely due to the fact that with this concept one has for the first time succeeded in giving an absolute definition of an interesting epistemological notion, i.e., one not depending on the formalism chosen. In all other cases treated previously, such as demonstrability or definability, one has been

It is this formalism-independence (as well as the persuasiveness of Turing's formulation) that has lent such strong weight to the Church-Turing thesis, to the effect that this mathematical definition does in fact correspond to our intuitive notion of computation. For whatever reasons , this persuasiveness has extended further, to the point where the Turing machine has been put forward as a model for the human mind, and mental processes equated with computation. Our own view of the matter is that such formalism-independence or "universality" is not to be expected in the setting of actual embedded systems with emergent functionality, and without pre-set "problems", let alone the necessity to act as universal computational devices.

Moreover, although the definition of Turing machine is extremely simple and is easily stated without an elaborate formalism, it leads to a variety of problems that would seem, if anything, to diminish the adaptive fitness of an organism encumbered with them: to the problematics of undecidability and unpredictability; to the distinction between the pure algorithmic (recursiveness) and the general manipulation of symbols (which is only recursively enumerable and so, a priori unpredictable, although one has a finite set of rules); and to unforeseen although simple results, (for example undecidability essentially can only occur when one allows erasing of symbols, etc...). Of course, Turing's purpose in devising these "machines" was precisely to investigate issues of this sort. But the perspective of the organism would be different.

(Although we shall not elaborate on game theory in this paper, we feel that similar arguments can be made against the relevance for real-world decision-making of the kind of "universal" rationality associated with game-theoretic pictures. We recall that in von Neumann's later work on game theory, he was very much influenced by the conception of the universal Turing machine. In the same sense as a universal Turing machine can simulate the behavior of any other Turing machine, game-theoretic strategies were intended to simulate the behavior of any opponent, and thus serve as a kind of general theory of rationality, of a rule-governed, machine-like character.)[23]

b) von Neumann's "self-reproducing automata"

At the end of the 40's, von Neumann began the development of a theory of "automata". This was envisaged to be a systematic theory of the principles of organization of these systems. However, while the theory was to remain at a logico-mathematical rather than empirical level , its concern was with actual, physically-embedded systems, both natural and artificial, rather than with formal systems. This meant that questions associated with the embedding (not the "contingent" features, but the logical concomitants of the embedding) were to be incorporated *within* the theory, rather than be regarded as incidental features to be given ad hoc treatment.

---

able to define them relative to a given language, and for each individual language it is clear that the one thus obtained is not the one looked for. For the concept of computability however, although it is only a special kind of demonstrability or decidability the situation is different. By a kind of miracle it is not necessary to distinguish orders, and the diagonal procedure does not lead outside the defined notion...."

(See also, M. Davis, "Why Gödel didn't have Church's thesis", Information and Control 54, pp. 3-24, 1982; J.C. Webb, "Mechanism, mentalism, and metamathematics", Reidel, 1980.)

23 See Philip Mirowski, "What were von Neumann and Morgenstern trying to accomplish ?", pp.113-148 in E. R. Weintraub, ed., "Toward a history of game theory", Duke U. Press, 1992.

One part of Von Neumann's work dealt with a theory of "self-reproducing" automata. This work, aside from its intrinsic merits, sheds light on possible interpretations that one may give to the notion of "mathematical theory" in the setting of embedded systems. One interpretation is that of "theory schema". In the context of self-reproducing automata, von Neumann constructed a schema requiring only a handful of components:(i) An automaton A, which when furnished with the description of any other automaton, in the form of an instruction I, will construct the automaton so described. All the automata constructed by A, as well as A itself, are presumed to have a place where an instruction can be inserted. (ii) An automaton B that makes a copy of any instruction I that is furnished to it. (iii) A control mechanism C that does the following: When A is furnished with an instruction I, the mechanism C will first cause A to construct the automaton described by I. Second, C will cause B to copy the instruction I and insert the copy I' in the automaton just constructed. Finally, C will separate the resulting entity from the system A+B+C and set it loose. (iv) Let D denote the system A+B+C, and let I_D be the instruction describing D. Insert I_D into A, and let E be the resulting entity. It can be verified that E is self-reproducing. This schema, developed several years before the work of Crick and Watson on the double-helix structure of DNA, predicted a fundamental role for a doubling, or copying mechanism. Even this degree of verisimilitude is quite remarkable, given the total absence of any empirics in the theory. As it stands, this theory has no way of even referring to, let alone predicting, specific biochemical properties of a cell.

Von Neumann wished to go beyond the level of schema, to some form of "realization" (though not necessarily a "physical implementation"). In some sense, the realization of such a schema might be viewed as the analog, in a quasi-embedded setting, of a model (in the standard model-theoretic sense) of a formal theory, but taking some account (real or simulated) of physical constraints. His initial plan was to construct a "kinematic model", incorporating geometric-kinematic aspects, such as movement, contact, positioning, cutting, etc., but ignoring problems of force and energy. Among the primitive elements of the model were to be logical and memory elements, sensing elements and effector elements, and "girders", to provide structural rigidity. This program was not actually carried out, it being not readily amenable to a strictly logico-mathematical treatment. Instead he constructed a "cellular automaton model", with less of a genuinely embedded character. This involved an infinite 2-dimensional array, with each cell containing a copy of the same 29-state finite automaton, communicating with its nearest neighbors.[24]

c) Formalization of linguistics

A good contemporary example is provided by attempts at development of theoretical linguistics. At least since the initial work of Chomsky in the early 1950's on generative grammar, the field has been in a continual state of ferment and reappraisal, giving rise to a succession of distinct contemporaneous "schools" of thought (several of them linked to the generative tradition, and sometimes centered around the same

---

24 There were also three other types of "model" contemplated but not actually constructed. A detailed account of von Neumann's work on automata, together with supplementary references, may be found in J. von Neumann, "Theory of self-reproducing automata", edited and completed by A.W, Burks, U. of Illinois Press, 1966.

individuals), differing not only on properly "technical" issues, but on fundamentals. These foundational concerns include, e.g., whether linguistics should be a study of language "in the abstract", or of a particular human biological endowment (cf. the different conceptions of Montague and of Chomsky of the notion of "universal grammar"); or, granting the latter viewpoint, how prominently the actual biology needs to figure (i.e., is it enough that the subject be "biologically predicated", or must it be "biologically based", and tied to specifics of neurobiology); whether syntax can or should be studied autonomously, or whether it must be treated as subordinate to "semantics" and/or "conceptual structure"), etc... This has lead to widely divergent viewpoints not only as to what kind of mathematical formalisms may be needed, but as to what formal elements must be incorporated or excluded, and even what the theory should be called upon to explain.[25]

Aside from its interest in its own right, we think that the work on theoretical linguistics may be relevant to the development of a theory for adaptive autonomous systems. For one thing, with its emphasis on heterogeneity and on the fundamental role of constraints, it provides a useful counterpoint to dynamical-system inspired pictures. Second, we think there is a significant parallel between the dichotomy of formal language vs. natural language and the dichotomy of formal system vs. natural system (or organism). In particular, we anticipate that the kind of theoretical framework required for the integrative organization of an organism will bear a qualitative resemblance to that needed for natural language. For example, the coordination of behaviors of an organism (with its qualities of "emergence", situatedness, "blurred specificity", etc...) is much closer in character to the grammar of a natural language than it is to the transition matrix of a finite automaton.[26]

---

25 Since we are not ourselves linguists, we shall not attempt to elaborate here on these foundational and methodological questions, nor shall we attempt to provide a representative set of references to the literature. We do, however, wish to mention, in passing, references to two approaches at rather different positions on the linguistics "spectrum": A presentation of the philosophy underlying generative grammar is given in N. Chomsky, "Knowledge of language: its nature, origin, and use", Praeger, 1986. The program of "cognitive grammar", which aims at a very comprehensive theory not in the generative grammar tradition, is described in R. W. Langacker, "Foundations of Cognitive Grammar", Vol . I, 1987; Vol. II, 1991, Stanford U. Press.

We also wish to mention an interesting treatment addressing the "bootstrapping" role that nonlinguistic (indeed, emotionally-grounded) cognitive faculties may need to play in supporting more properly language-specialized "modules", and in precipitating the child's path to "language acquisition". (See John L. Locke, "The child's path to spoken language", Harvard U. Press, 1993).

26 This kind of analogy has been made before from a number of distinct (and even conflicting) perspectives. For example, Lashley, at the 1948 Hixon Symposium (at which von Neumann presented his schema for self-reproducing automata) speaks of the problem of "syntax of action" (see K.S. Lashley, "The problem of serial order in behavior", in L.A. Jeffress, ed., "Cerebral mechanisms in behavior", Wiley &Sons, 1951); Miller, Galanter, and Pribram (see "Plans and the structure of behavior", 1960) propose an analogy between "plans" and phrase-markers, and ask for the analogue of a transformational grammar in this setting; more recently, some investigators have argued for the evolution of primate motor systems as preadaptations for the development of human language ability, with a consequent parallelism between temporal organization of motor activity and of language (see, for example, J. Kien, "Developments in the pongid and human motor systems as preadaptations for the evolution of human language ability", to appear in Studies of Language Origins).

3.8 What kind of mathematical approach do we envision ?

a) Conceptual architectures

We regard our primary problem to be the development of "logical" or "conceptual" architectures for adaptive autonomous systems. From this perspective, all questions of technical feasibility aside, it would be inappropriate to focus effort on an actual full-scale implementation of a "total", top-to-bottom integrative treatment of a chosen model system.[27] Rather, for the purpose at hand it is more relevant to "integrate" a model system in principle than in practice. That is, to use the model system as a source of gedanken experiments to bring into focus the "necessary and sufficient" conceptual components, and as a touchstone against which to compare our attempts at a synthesis of these components. So as not to inadvertently omit "components" essential for "self-containedness"[28], it will be important to maintain a scrupulous distinction between what the system can actually do autonomously, given the architecture available to it, and what in fact is an act of "transference" on the part of the theorist. In particular, it will be important to distinguish the "semantics" and "interpretation" internal to the system itself from attributions of meaning made by an external "observer" (even one who is not limited to observing external behavior, but who can look inside the system). At a minimum, we expect that such an effort at development of conceptual architectures should result in a clarification of the vocabulary. This in itself is a far from easy task because of the divergence and disagreement about the intuitive meaning of many terms used in artificial intelligence, in theoretical computer science, and in neurobiology.

b) Choice of model system

Even at the level of "conceptual architecture" we shall not attempt to incorporate, or study, all desirable functionalities at once, but shall proceed incrementally. Thus, we shall be experimenting with a succession of conceptual architectures. These may be overlapping to a greater or lesser extent, but need not be nested, since we may be using distinct architectures to highlight distinct sets of issues. For this reason it may be preferable to make use of several model systems, each particularly suited to a specific range of questions (either because of inherent properties or capacities of the system, or because of the types of data or experimental techniques available), rather than insisting on a single model system. By the same token, while our focus is on principles of integrative system organization, our models need not be limited to "entire" organisms.

c) Plausibility

We have no hesitation as regards incorporating exploratory or provisional ideas in our "models". Indeed, we don't see how this can be avoided. However, we do want to observe at least some minimal guidelines as to "plausibility". If we are focusing on

---

27 Such a full-scale treatment was proposed for the model system *C. elegans* in the Working Paper cited in Chapter 1.

28 The notion of "self-containedness" (relative to an embedding) can perhaps be regarded as a kind of analogue, in the setting of embedded (or embodied) systems, of the notion of "closure" or "completeness" in the context of formal systems. We are aware that the notion is a highly problematic one from the standpoint of "conceptual architectures", since certain of the necessary "logical" components may be "entailed" not by other logical components, but rather by the specifics of the supporting physical substrate. (See also the remarks in Section 2.14 and in Section 3.5(ii)).

biological organization (resp., artificial systems), we want to respect the actual possibilities of biology (resp., technology). By this we do not mean, for example, that a mathematical theory must take full account of the details of biochemical processes. We mean simply that the functionalities or structures that we introduce should not be at variance with known biological principles or modes of operation.

d) Maintaining coherence: "gluing together" local logics

As emphasized both in the present chapter and in the preceding one, we want to give center stage to the issue of "coherence". The general idea is that an adaptive autonomous system is embedded in a rich environment, and develops "local" logics both for analyzing and acting on its environment, and to satisfy its various goals. The problem is that these local logics must coexist and must, in some sense, be "glued together" so as to be compatible. Our concern will be to elucidate the nature of this "gluing". As yet there is no standard mathematical theory that is directly applicable, only a number of suggestive mathematical analogies to go by. For example, this idea is not totally unlike the idea of a "variety" or a "scheme", in geometry, which is obtained by pasting together "small" building blocks (like balls in differential geometry, or affine Zariski open sets in algebraic geometry).[29]

Here "logic" should not be taken in its usual mathematical sense, e.g., as Boolean or predicate logic, nor is "local" intended in a technical sense. Rather the term "local logic" is meant to indicate the "coherence" of various subsystems in relation to their respective environments. The problem is to maintain coherence for the system as a whole. (As noted in Section 2.14, the pitfall for embedded systems is incoherence rather than logical inconsistency, as would be the case for formal systems).

We use the term "gluing" in a broad sense, to indicate a kind of "non-independence" or " imposed coordination" among the entities "glued". In particular, it is in no way limited to situations where there is an actual physical joining or interconnection of "parts". We shall illustrate what we have in mind with a couple of schematic examples in which subsystems are represented as abstract control "modules" (please refer to Figure 3.1). We begin with a "parallel juxtaposition" as in diagram (i), in which the modules **A** and **B** have access to distinct sets of effectors (or sensors). In diagram (ii) the situation has been altered so that **A** and **B** have access to the same set of effectors, resulting in a "juxtaposition gluing". This gluing is, of itself, unsatisfactory in general, because the lack of coordination between **A** and **B** can give rise to conflicting "demands" on the effectors. Two ways of introducing such coordination via additional "gluing" are illustrated in diagrams (iii) and (iv), respectively. The first involves the direct interaction of **A** and **B**, say via reciprocal inhibition. The second involves the introduction of an additional set of modules **C** controlling the "communication" between **A** and **B** and the effectors. In diagram (iv) we have included not only "forward" arrows from **C** to **A** and **B**, but also reciprocal, "backward" arrows from **A** and **B** to **C**. This is not only so that "information can be passed back to **C**. In addition (cf. Section 6.10c)), we

---

29 An extensive discussion of questions (and mathematical analogies) associated with the "gluing together" of entities of heterogeneous type is presented in Section 3.3 of the *C. elegans* Working Paper cited in Chapter 1, from a perspective close to that of the present discussion.

want to allow for the possibility that **A** and **B** may learn to access **C**, and to "ask for help".[30]

e) Dynamic vs. static gluing

Returning to the above analogy with geometric gluing, we wish to emphasize one major conceptual difference (among several) between the concept of gluing in geometry and the concept of gluing local logics. In the geometric setting the gluing is fixed, or static (though perhaps a parametrized family of such glued structures may be considered concurrently). In the present context of "local logics" the gluing will of necessity be dynamic, or time-varying. This stems from the fact that we are dealing with an adaptive system, which implies that its architecture should be able to modify itself depending on the interaction of the system with the external environment. In other words, even leaving out of consideration strictly "developmental" or maturation processes, a particular gluing that is consonant with the system's being well adapted to one class of environmental situations may become inadequate or infelicitous if the environmental situation changes. To some extent this shift from adaptedness to nonadaptedness may be linked to differences with respect to certain of the statistical regularities presented by the distinct classes of environment.[31]

In addition to adaptation to the external environment, there may be a need for architectural modification or reorganization in response to changes in the internal milieu, the most drastic being "breakdown" or "failure" of system components. (Such damage certainly occurs in biological systems, including nervous systems, as well as in artificial systems, with the major difference that the former are much less brittle and far more failure-tolerant. This is reflected in the corresponding modes of adjustment. In biological systems this adjustment is based on a general context-dependency of over-all system functionality, rather than on "redundancy". It is more akin to a process of "relaxation" from one coherent dynamic functional equilibrium to another than to the shunting of a sharply "pre-specified" functionality to a "back-up" subsystem.)

---

30 In our broad use of the term "gluing", the "convergence zones" of A. Damasio's retroactivation schema (cf. the footnote to Section 7.12(iii)), which mediate the synchronous reactivation of neuronal firing patterns in lower-order convergence zones or in primary cortices, may be viewed as a kind of gluing of these lower-order zones (or primary cortices). Notice that in this setting the same lower-order zones (including the primary cortices) may participate in a multiplicity of such gluings.

31 We should perhaps clarify that, while we use the term "gluing" in a broad sense, we don't wish to deprive it of all meaning by using it in an all-encompassing sense. Thus, we will not wish to regard every alteration in system architecture as ipso facto a change in "gluing".

4. Descriptions and prescriptions for adaptive systems

In this chapter we begin to discuss certain "primitive" intentional concepts which are typically attached to intuitive notions of intelligence and adaptation, and are used for system description and prescription.[1] These concepts are also used freely in artificial intelligence and robotics. Our aim here is not to set up formal definitions, but rather to do some clarification of the intuitive content of these concepts, as a preliminary to eventually developing a more adequate technical vocabulary. We shall focus on the two notions of "constraints" and "specifications", and emphasize certain basic distinctions between them. As it stands, the notion of "specification" while probably indispensable for artificial systems, carries certain connotations which make it rather ill-suited for the description of natural systems, especially adaptive autonomous systems. However, it would be desirable to have some sort of "substitute" notion which is adapted to the context of natural systems. We shall examine certain respects in which this substitute will need to differ from the original notion.

A. Constraints and specifications

4.1 Necessity vs. choice

The basic distinction which it will be useful to make between "constraints" and "specifications" centers on the polarity between "necessity" and "choice". Thus, we shall use the term "constraint" as a generic name for relations which are *necessarily* satisfied by a given system. Here, the system might be the adaptive system that we are trying to study, a certain subsystem of this adaptive system, the environment considered as a system, the total system which is the union of the adaptive system in its environment etc... The term "specification", by contrast, will refer to relations that may be "imposed" or "chosen", but need not hold a priori. (In speaking of "system" we have in mind embedded vs. formal systems, and our discussion is primarily intended for this context. However, some of the points made may also be germane to the setting of formal systems).

We shall elaborate in the sections below on what we have in mind when we speak of "necessity" or "choice". Before doing so, we want to make a few clarifications. (i) Although we feel it is important to emphasize the basic distinction between the two notions, in any actual context the "constraints" and "specifications" are highly intertwined; for example, constraints may give rise to specifications. which may give rise to additional constraints, etc... (ii) Constraints may play an enabling as well as a limiting role. For example, specifications may be dependent on constraints (such as physical laws) for their realization. Similarly, constraints may serve as mediators between specifications of structure or connectivity of the system and specifications of its behavior. As another example, the very restrictions of choice imposed by constraints may, in fact, facilitate the process of choosing. (iii) We do not view the distinction between "constraints" and specifications" as a matter of fundamental "ontology", but rather as a useful perspective

---

1 This is not to say that they need to form the "primitives" of an actual theory.

to bring to a study of system organization. Certainly there are cases where it is difficult to determine which term (if either) is appropriate.

4.2 Constraints

a) Necessity

When we speak of relations which are necessarily satisfied, "necessity" can signify either logical necessity or physical necessity. For example, any physical or chemical law is a physical constraint; a logical consequence of a certain set of "hypotheses" (regarding the system and/or its environment) is a logical constraint; the fact that shared variables take the same value is a logical constraint, etc...[2] We also wish to allow "material consequences" as well as logical consequences. "Hypotheses" can take a variety of forms, for example: (i) Other constraints; (ii) Particular properties that happen, in fact, to hold for the system and/or its environment. Sometimes it will be natural to speak of these facts as themselves constituting "constraints", e.g., the limited availability of certain resources. (iii) Specifications. These need not be specifications that have in fact been implemented, but may only be under consideration for implementation. Hence, the term "hypothesis". The latitude allowed to the term "hypothesis" reflects the relative, or conditional, or context-dependent character we wish to permit to the term "necessity".[3]

b) Some basic constraints

Let us now examine certain basic constraints associated with an adaptive system which is physically (and perhaps biologically) embedded in its environment. These are, by and large, consequences or concomitants of the "fundamental specification" (see Section 4.4a)).:

(i) Homeostatic constraints
(ii) Sensory constraints
(iii) Effector constraints
(iv) Chronological constraints
(v) Constraints on the internal structure of the system

c) Homeostatic constraints

We include under this heading the basic regulatory constraints, such as those associated with the traditional "cybernetics" perspective. Notable among these are the "metabolic constraints", these being associated to the fact that any realistic organism (indeed, any system, whether natural or artificial) needs energy to function (and in general, several different sources of energy), produces entropy (and so degrades its environment), and in general can only function when certain physical or chemical variables, either of the environment or of the system itself are maintained between certain

---

2 We are not attempting here to set up some kind of basic taxonomy of constraint types; the labels "physical", "logical", etc. are simply used here for convenience of reference.

3 Certainly we are using the terms "conditional" or "context-dependent" in an intuitive sense here. However, it is possible that some type of formalization, akin to the formalization of "conditioning" in the stochastic process context (conditional probabilites; conditional expectations) may be appropriate.

bounds (which may or may not be sharp)[4]. For example, a living organism cannot live in too high or too low temperature conditions. Among the metabolic constraints is the fact that the organism's energy reserves are continually being depleted, even if the organism "does nothing" (i.e., engages in no "activities" other than those minimally necessary to stay alive, such as respiration). This implies that in order to maintain the relevant variables within the admissible bounds, the organism must take action within its environment (to acquire food).[5] Provisionally, we shall also subsume processes of reproduction under the heading of "homeostatic constraints".

d) Sensory constraints

We have already discussed (see Section 3.3 c)) the inherently emergent character of the system's categorization processes. Here we shall allow ourselves to speak more loosely, and to reify "information", as if it were something to be "extracted" from the environment. From this perspective, any adaptive system (and in fact any natural or artificial system) has access to its environment through perceptual devices which extract certain information from the environment. This information is then "shared" by various subsystems of the system. The implications arising from this need to share form the basic class of sensory constraints.

The second class of sensory constraints corresponds to the fact that an adaptive system has few perceptual resources or modalities with which to interface with a rich environment. Given these constraints, the system must find strategies to overcome the "information bottleneck".[6]

---

4 The satisfaction of the metabolic constraints may, via "material implication", involve the broader class of homeostatic constraints, not concerned exclusively with energy. For example, there may be a need for support processes to transport "materials" which are used at other "stations" to actually produce energy. This starts to look like a "factory".

5 Actually, we should be a bit more careful here and distinguish, e.g., between metabolic *constraints* and metabolic *specifications* . Thus, we would continue to regard as metabolic *constraints* physical or chemical laws, say that ingestion of food produces a rise in certain metabolic variables and that a decay of the external temperature induces a decay of the internal temperature or a consumption of a certain amount of metabolic resources. Another metabolic *constraint* is that certain metabolic variables are always decreasing, with the rate of decrease depending on the animal's activities, but always strictly $\neq 0$. From these constraints and the fundamental specification would follow the metabolic *specification* that the system's internal metabolic variables be maintained within a certain domain. Of course, this and, consequently, the "fundamental specification"(see Section 4.4a)) may, in fact, fail to be fulfilled, since the animal can starve or freeze (or die of "natural causes"). However, under this interpretation only *specifications* and not *constraints* are being violated.

These specifications do not, however, of themselves imply precisely which actions the organism must take at any particular time or in any particular context in order to acquire food. This may, to a greater or lesser extent, permit "choice" on the part of the organism. (We note in passing one basic difference between this continual need for energy on the part of an organism and the energy needs of a machine, say a computer. The machine can be turned off when not in use , so that it consumes no energy. On the other hand, if the external temperature is allowed to become sufficiently low, the machine may be damaged whether it is on or off at the time.)

6 One is tempted here to emphasize the often-made parallel between energy (in this case, the metabolic constraints) and information (in this case, the sensory constraints). However, even if one goes beyond the picture of "extracting" information from the environment to that of "constructing" information from "raw materials" drawn from the environment, the supposed parallel is highly misleading: For example: it tends to confound information in the sense of Shannon with "semantic" information; it suggests an interconvertability of different "forms" of information; and it, in particular, suggests that information is in some sense a scalar quantity.

e) Effector constraints

Any adaptive system acts on its environment (as, for that matter, do essentially all natural or artificial systems)[7]. The actions are performed by means of various effector devices. Distinct subsystems of the system may have joint access to certain of these effectors, and thus are obliged to share the possibilities for action. This is the basic class of action constraints.

The second class of effector constraints correspond to the fact that an adaptive system has few effector resources with which to act on its environment, and with which to carry out a rich repertoire of behaviors. Again, given these constraints, the system must find strategies to overcome the "action bottleneck".

Notice, here, the symmetry between the sensory constraints and the effector constraints. We note also that considerations similar to those we have discussed in connection with the overall system embedded in its external environment hold as well for subsystems embedded in their respective "proper" environments (as defined in Section 7.10a)).[8]

f) Chronological constraints

This refers to the various constraints associated with time, which is itself a "shared" variable for the environment and for the various subsystems constituting the system, resulting in the necessity for the temporal coordination of events and actions. Among the constraints are those associated with temporal coincidence, temporal duration, and temporal ordering. In particular the system needs to work in "real time". (See also the footnote to Section 7.5(i)).

g) Constraints on the internal structure of the system

This is not to suggest that this internal organization is either rigid or static, but rather that the "tools" or "functionalities" available at any time are limited in number and/or "capacity" (an example being limitations of various forms of "memory capacity").[9]

4.3 Specifications for an artificial system

Let us for the moment restrict ourselves to the class of artificial systems (i.e., man-made systems, adaptive or not). As a concomitant of being embedded, these systems must ipso facto satisfy certain constraints with respect to their environment, including (at

---

7 This action may be incidental to the primary functionality of the system, as in the case of the heat generated by a computer.

8 We should perhaps clarify that, in focusing attention here on sensors and effectors, we are not claiming that they somehow come "first", either temporally or logically, with the internal architecture "filled in" secondarily.

9 We note in passing that the sensory and effector constraints could perhaps be included under this heading since they are linked to a scarcity of "means" (either sensors or effectors) internal to the system (as well as to an "overabundance" external to the system). This may be contrasted with the metabolic constraints. These constraints are partly associated with the internal constitution of the organism, but also (at least as regards the modes of behavior by which the system can cope with these constraints) are partly attributable to a scarcity of external resources.

least in the case of adaptive autonomous systems) the five classes of basic constraints enumerated in the preceding section.

a) Specifications vs. constraints

The designer[10] of an artificial system always designs with certain purposes in mind. We shall reserve the term "specifications" for the various prescriptions imposed on the system by the designer. This designation is to distinguish these prescriptions from the constraints, or necessary relations satisfied by the system. While the constraints are necessarily fulfilled, regardless of what the specifications happen to be, the specifications may not be fulfilled, in which case the system "fails" in some sense. (We shall return to the failure problem in Section 4B).[11] An additional sense in which the specifications are not "necessary" relations is that the designer could have imposed a different set of specifications.

Some examples of specifications are prescriptions regarding: structural components, connectivity or interface relations, types of information used by the system, behavior types, preference relations or choice procedures among behaviors or resource allocation, optimality criteria, goals or aims that the system is to achieve, how the system is to be assembled, etc... (There may also be guidelines, or principles of evaluation, that the designer uses for selecting among different qualitative classes of design specifications, e.g., emphasizing parallelism, or modularity, or reusability of components, etc...).

b) Descriptions and prescriptions

In "writing out" the specifications of the system, the designer may make use of various modes or formats of expression, including a variety of formal specification languages.[12] However, at least when the specification is for an embedded system, there will typically be no need to attempt to formally write out the constraints on the system as part of the system prescription. The constraints come along "for free", even though they may be doing most of the "work" of enforcing the intent of the specifications. (For example, if the system were to be embedded in a different environment, with different constraints than those holding in the original environment, the original system specification may no longer be appropriate). However, for certain purposes, such as system simulations, it may be necessary to explicitly include certain of the constraints as part of the system description. (See Section 4B).

---

10 We recognize that to speak of "the" designer is an idealization. Most design projects of even a "moderate" scale will involve teams of designers, even at the level of the "conceptual design". Achieving global coherence of the resulting design given the distributed nature of the design process is a major problem in its own right.

11 We are speaking here as if there is an already constructed system, on which the designer is seeking to impose some additional specifications, say regarding behavior. On the other hand, the unfulfilled specifications may be a design "on paper" which cannot be carried through to the actual construction stage. This may stem from a variety of possible difficulties, including: (i) The specifications are inherently inconsistent or incoherent. (ii) The resources for construction are insufficient. (iv) The specifications are insufficiently complete. (v) The specifications are in a "declarative" form, and there is no procedure available for converting them to the "imperative" mode.

12 A word of caution. In the context of these formal languages it is not uncommon for specifications (in our sense) to be referred to as "constraints". The same holds for the *C. elegans* Working Paper cited in Chapter 1.

We have spoken above of "writing out" specifications but, as we shall discuss later (see Sections 5B and 7.11a)) there are cases (indeed, we regard them as the most interesting cases) where it is not possible to do so explicitly. By "explicit" we refer not only to extensional descriptions, but also finite intensional descriptions which might "generate" infinite extensional descriptions.[13] Of course, the effects of "specifications" include not only the specifications themselves but the ensuing logical and material consequences.[14] These propagated effects, together with the fact that the specifications may not be explicitly describable, can contribute to the "unpredictability" of the system.

4.4 Specifications for a natural system

a) "Specifications" in a wider sense

The term "specifications" tends to carry connotations of a "finalist" point of view. Hence, while the notion of "constraint" does not need reinterpretation in the context of natural systems , the term "specification" will need to be taken in an analogical sense, because there is no designer a priori.[15]

There are at least three analogical senses in which one may wish to speak of "specifications" in this context: (i) Evolution is sometimes metaphorically spoken of as the "designer" (or tinkerer) of living organisms.[16] (ii) An external observer may observe

---

13 This point bears some relation to the observation made by von Neumann, in connection with the question of trying to give a description of "visual analogy". He proposed that the simplest way to do this might be to describe the *structure* of a device which makes such analogies. He expressed this alternatively as the conjecture that in such settings a real object might constitute the simplest description of itself. He also suggested that "some results in modern logic" supported this conjecture, and refererred to a result of Gödel.. (See J. von Neumann, "Theory of self-reproducing automata", edited and completed by A.W, Burks, U. of Illinois Press, 1966, esp. pp. 53-56. As described in these pages, Burks, the editor of this posthumous work, consulted Gödel to determine what results in logic von Neumann might have been referring to. Gödel suggested two possibilities: " I think that the theorem of mine which von Neumann refers to is not that on the existence of undecidable propositions or that on the lengths of proofs but rather the fact that the complete epistemological description of a [formal] language A cannot be given in the same language, because the concept of truth of sentences of A cannot be defined in A. It is this theorem which is the true reason fot the existence of undecidable propositions in the formal systems containing arithmetic....The same theorem was proved by Tarski in his paper on the concept of truth... Now this theorem shows that the description of what a mechanism is doing in certain cases is more involved than the description of the mechanism, in the sense that it requires new and more abstract primitive terms, namely higher types. However, this implies nothing as to the number of symbols necessary, where the relationship may very well be in the opposite direction... However, what von Neumann perhaps had in mind appears more clearly from the universal Turing machine. There it might be said that the complete description of its behavior is infinite because, in view of the non-existence of a decision procedure predicting its behavior, the complete description could be given only by an enumeration of all instances. *Of course this presupposes that only decidable descriptions are considered to be complete descriptions*, but this is in line with the finitistic way of thinking...[emphasis ours]). As regards "types", see also Section 7.3. As regards "predictability", see Section 3.6(ii).

14 It is somewhat problematical as to whether these effects should be classed as "constraints" or (implicit) "specifications" (or neither). They are perhaps best regarded as "closure" and compatibility conditions on the overall set of specifications.

15 We are not altogether happy with speaking of "specifications" outside the artificial system setting, and regard it as a purely provisional terminology which we are using in order to highlight certain distinctions between artificial and natural systems.

16 In this setting, people have in fact begun to speak in terms of the interplay, or creative tension, between the inherent freedom of specification allowed to the "designer" and the underlying constraints.

the system and, on the basis of the observations made, may conclude that the system satisfies certain possible relations which are non-necessary. The observer, in order to "rationalize" and to predict the behavior of the system, may impute some sort of underlying intentionality, or self-imposed behavioral "specifications" to the system. This is the so-called "intentional stance" of Dennett. (iii) An autonomous system may, in fact, impose "specifications" on itself, as regards goals, plans, behaviors, modes of effecting these behaviors, etc... This imposition of specifications may be only partially, if at all, "volitional", and may incorporate both "conscious" and "unconscious" processes. These "specifications" will, however, be markedly different in kind from the "sharp" specifications typically imposed in the context of artificial systems; instead they will have a "blurred" or "vague" character (cf. paragraph b) below, and the discussion of "vague specifications" in Sections 2.10, 7.11a), and 7.13c)). One additional distinction (to be developed more fully in Section 4B) is that in the case of natural systems, the specifications imposed have relevance to the system itself[17]; hence failure to satisfy these specifications (most notably, the "fundamental specification") has consequences for the system itself.

So we can widen the scope of the notion of "specifications" to include possible relations which are "in general satisfied", but are not "necessarily satisfied" (i.e., they are not constraints), and which are either imposed on the system by a "designer" (in the wide sense), or imputed to the system by an external observer, or imposed on the system by itself.

One class of "specifications" deserves special mention, namely those associated with the proto-goal of autonomous systems, namely to maintain themselves as systems (cf. Section 3.3a)). In the case of a biological system this implies, among other things, that the system must stay alive. We shall speak of this proto-goal as itself being a "specification" (indeed, the fundamental specification). Certainly, at least for biological systems, this specification is eventually violated.[18] (We note that this is a "fundamental" specification, in the sense of being a sine qua non, for all systems, whether or not autonomous. However, in the case of autonomous systems this is the underlying rationale of the system organization, and the foundation of the "value system", whereas in the case of a (non-autonomous) artificial system it is no more than an implicit, though critical, side-condition on the design. Also, in the case of autonomous systems, this is the "responsibility" of the system itself during "somatic time").

b) Vague specifications

At present we are not prepared to enter into a genuinely technical development, and wish here only to add a couple of comments to the intuitive discussion of this notion

---

(See, e.g., S.A. Kauffman, "The origins of order: self-organization and selection in evolution", Oxford U. Press, 1993; S.J. Gould, "The evolutionary biology of constraint", Daedalus, Spring 1980, pp. 39-52.).

In any case one would not speak here of "goals" or "external (to the organism) evaluation criteria" of the "designer". The perspective is ex post facto "internalist".

17 Even if they are "imposed" by evolution.

18 It is perhaps the case that, from an internalist perspective, this specification is not, in fact, violated. From this perspective, "time" begins with the coming into being of the organism and comes to an end with the death of the organism. (This also suggests the possible alternative route of regarding this proto-goal as the fundamental *constraint* rather than the fundamental *specification*).

given elsewhere in this paper (cf. Sections 2.10, 7.11a), and 7.13c)). In the preceding chapters we have stressed fundamental differences in character between artificial systems (as currently conceived) and natural systems (or, more generally, adaptive autonomous systems). In our view, a key to coming to grips with these differences, at a technical level, lies with the elucidation of the differences in the corresponding notions of "specification". These differences include: vagueness vs. sharpness (not to be confused with partiality vs. completeness), situatedness vs. context-independence, emergence vs. external imposition, and self-relevance vs. outward-directedness. As regards situatedness, we want to mention the importance of accommodating incrementality of specification, including allowing multiple time-scales for "choice". This includes not only the broad distinction between choices made on evolutionary vs. developmental vs. somatic time-scales, but also different time-frames within somatic time (cf. Section 2.15). This, in particular, allows for the dependence of certain "decisions" on data, input, experiences, etc. which could not have been anticipated prior to the actual experience, as well as mitigating the "computational burden" associated with the perfect foresight and understanding assumed in game-theoretic "rationality" or in "classical planning" in the sense of AI (cf. Section 3.7a)).[19] As regards the choices made on different time scales, one should remember that among the "choices" or "specifications" made, say, on an evolutionary time scale are some which, in effect, determine "internal resources" (i.e., the system's "architecture", including its nervous system if it has one, its sensory and motor apparatus, etc.) available to the system for making choices on a developmental or somatic time scale, including decision-making "biases" or "biasing mechanisms" (cf. Section 3.3d)).[20]

It is not sufficient to "declare" specifications, but to actually put them into effect. This is an issue in both the artificial and natural system contexts. It involves, in particular, converting or refining "high-level" specifications, perhaps via a succession of stages[21], into explicitly executable "low-level" specifications. The way this process is conceived is, however, very different in the two distinct contexts (cf. the remarks on "context-dependent compilation" in Section 7.13c)). In the artificial system context, one emphasizes the top-down character of the refinement process. The high-level specification is taken, by and large, as a given, and the issue is to verify (and perhaps to "prove", using formal verification methods) that the low-level implementation does, in fact, faithfully instantiate the high-level specification. We envision the process to be quite different in the setting of natural systems (or adaptive autonomous systems). The high-level "specification" is vague rather than sharp, and the process of "refinement" has more the character of actively "deciding", based on the "local" context, what meaning should be given to the specification than of passively accepting a set of instructions. (As we have

---

19 We feel that the notion of "vague specification" is also more apposite in the context of development (of an organism) than is the notion of "specification" or "instruction" in the traditional sense. We have in mind, in particular, the role of the genome in development.

20 Since an autonomous system to some extent chooses its own "problems", based on its over-all "architecture", these "specifications" not only influence the somatic-time decision-making processes, but also the "problems" around which decisions are made.

21 As will be emphasized in Chapter 7, such notions of "stages" or "levels" should not be taken too literally in the autonomous system context. One can envisage many types of "organizational chart" drawn from many different perspectives. We expect "hierarchy" to not to figure prominently in most of these charts

already emphasized in Section 2.10, we do not regard this "vagueness" as a weakness, but as a key factor sustaining system autonomy and adaptiveness. We shall argue in Section 4B that this vagueness and the associated differences in the specification "refinement" process, point to a fundamental non-algorithmicity of operation). Moreover, higher-level specifications (even the "highest-level")[22] are subject to revision in conjunction with this refinement process. (As one example, events occurring within one somatic time-frame may precipitate revision of "plans" associated with longer time-frames; cf. the remarks on "expectation" in Section 7.9d)). In particular, the processes of construction (or selection) of specifications and the processes of implementation are intertwined rather than consecutive. This intertwining contributes towards assuring "coherence" of behavior (and of specifications). We feel that the above processes cannot adequately be subsumed under the rubric of "feedback loops".[23]

c) Natural systems as dynamic vs. static

By this, we refer not to "dynamical systems" in the usual sense of the term, but to genuinely emergent properties of the system, including structural and functional reorganization or alteration, including gain, loss, bifurcation, or refinement of internal "resources" or functionalities. Such emergence, leading as it inevitably does, to "emergence relative to models" (cf. the footnote to Section 3.3c)), presents basic difficulties of principle to modeling and simulation studies. One manifestation of such difficulties is that the traditional "dynamical system" picture of "time evolution" is inadequate to accommodate such emergence. This amounts to treating time as just as additional spatial parameter. In particular, this picture presupposes an a priori fixed "state space", with the time evolution corresponding to moving from one point in this fixed state space to another (or, alternately, as assigning a new set of values to a fixed set of "slots").

In the context of artificial systems one may, to a greater or lesser extent be able to disregard issues of emergence, but we feel that this is not the case for natural systems (or adaptive autonomous systems). For example, a characteristic feature of such systems is the emergence of "macroscopic" organization, very possibly not "predictable" from the "microscopic" organization, yet which can causally influence or bias the future time evolution of the microscopic organization.[24] This kind of emergence, together with its

---

22 We reiterate here: Despite our reference to "levels", we do not have in mind anything like a hierarchical organization.

23 We feel that what is involved here is not a matter of feedback vs. feed-forward, or of closed-loop vs. open-loop control. Rather, what is pointed to is the need for a different conception of the nature of "control".

24 This alteration of the original microscopic dynamics by the newly emergent macroscopic structures is a fundamental difference between emergence in the above sense and "computational emergence" (cf. the footnote to Section 3.3c)), such as figures in the work on "artificial life". (See, e.g., C. G. Langton, ed., "Artificial Life", Santa Fe Institute Studies, Vol. VI, Addison-Wesley, 1989). In simulation studies such causal influence can be "put in by hand" by the modeler, via the introduction of new models, but this substitutes the cleverness of the modeler for that of the model.

One framework (that of "component" systems) for accommodating emergent systems has been proposed by George Kampis. (See, e.g., "Self-modifying systems in biology and cognitive science: a new framework for dynamics, information and complexity", Pergamon Press, 1991).

As regards predictability of the macroscopic organization (and of its causal consequences), compare the remarks in Section 3.6c)(ii).

attendant "unpredictability", is a prominent characteristic on both the developmental and somatic time scales, but it is, perhaps, taken most seriously in the setting of the "historical" sciences.[25]

We have already emphasized in paragraph b) that both the "situated-interpretation" and "revision" of specifications are ongoing processes in the setting of natural systems. On the one hand, this is necessary because of the above-noted dynamic character of the system itself. Reciprocally, the necessity of embodying and accommodating this "specification dynamics" is one of the "reasons" for the dynamic character of the system.

4.5 Examples of specifications and constraints

Example 1: optimal control theory

In this example, we have a dynamical system whose state depends on the value of a control parameter fixed by an external agent. In this case, one expresses the constraints by means of differential equations (giving the law of motion). The specifications are given by the usual boundary conditions and by the functional of the trajectory which should be optimized. The system may "fail" in the sense that there might be no trajectory satisfying the boundary conditions (i.e., these boundary conditions are not reachable).

Example 2 : a thermostat

Let us now consider an archetypal servo-control problem, namely the control of the temperature of a room by a thermostat. The constraints are given by: the heat conduction law; the physical structure of the thermostat and its response function to a variation of temperature; the fact that when a certain amount of fuel is burnt, the temperature of the room rises; the fact that when the outside temperature is varying, the room temperature has also to vary accordingly. The specification is that the room temperature should be maintained within two bounds T_1, T_2.

We can also conceive of a more complex thermostat which indicates the amount of fuel left in the tank. Then an additional specification might be that the level of fuel in the tank should be maintained within two bounds f_1, f_2. An additional relevant constraint would be the relation between an increase of temperature and the consumption of fuel. This example illustrates the conditional character of the "necessity" associated to

---

25 For discussion of "hierarchy theory" in the context of evolution, see, for example S.J. Gould, "Punctuated equilibrium in fact and theory", and N. Eldridge, "Punctuated equilibria, rates of change, and large-scale entities in evolutionary systems", both in A. Somit and S.A. Peterson, eds., "The dynamics of evolution", Cornell U. Press, 1992.

Philip Mirowski argues the need for other mathematical frameworks (based more on a hermeneutic perspective) than those of a mechanistic, mathematical-physics character in order to describe economic organization, in particular institutional economics. For example, he suggests: "Institutions can be understood as socially constructed invariants that provide the actors who participate in them with the means and resources to cope with change and diversity..." (See, p. 132 of P. Mirowski, "Against mechanism: protecting economics from science", Rowman & Littlefield, 1988).

The distinguished neuroscientist Roger Sperry has, over the years, espoused a framework within which "consciousness [is] a nonreductive emergent with causal potency and downward control ". (See p.204 of R.W. Sperry, "Mind-brain interaction: mentalism, yes; dualism no", Neuroscience Vol. 5, 1980, pp. 195-206.

constraints. Thus , if we change the design of the heating system (converting from the use of fuel oil to the use of coal or electricity) then the above constraint changes.

Example 3 : a railway station

Let us examine two examples in the context of a railway station. A railway station is an extremely large control system. An obvious physical constraint is that two trains cannot simultaneously occupy the same place and, as a result, an obviously critical safety specification is that trains should not collide. For added security , one may wish to impose the more stringent specification that two trains should not be on the same track at the same time. A different class of specifications, not directly concerned with safety, might center on the scheduling of the train service.[26]

Example 4 : an elevator

This is another example of a "transport system". Constraints would include the law of gravity, Newton's equation, cable-strength... A specification on the design of the elevator might include the maximum load and speed under which it will be expected to operate.[27] Here the "constraints" and "specifications" thus far considered imply additional consequences, for example, specifications as to the maximum power output required from the motor.

There are also specifications regarding "scheduling", i.e., the policies of response to user requests for service. One high-level specification might be that people should wait the smallest amount of time possible. Another might be that the elevator takes on waiting passengers according to some set priority (for example, first-come first-served). Still a third might be that the elevator does the minimal amount of travel between floors. It is clear that, in some sense, one would like to satisfy all these specifications together, but this may be impossible in principle. Thus, this example is a case where it is not at all obvious as to what high-level specification should be imposed. Moreover, having determined on a high-level specification, it may be a very difficult (if not intractable) problem to determine how to program the elevator so as to (approximately) conform to this specification, in particular if the elevator has limited computational resources and can only be allowed minimal time to reach a "decision". The implementation problem becomes considerably more difficult if the high-level specification is to be satisfied conjointly by a bank of elevators rather than a single elevator.

Example 5 : biological organisms

We have already discussed some basic constraints on an organism in Section 4.2, and have given examples of specifications made both on an evolutionary and on a somatic time-scale in Section 4.4. Some of these latter specifications will be fleshed out a bit more in Chapter 7.

---

26 Apparently, the study of the use of "formal specification methods" in connection with safety-critical applications (including railway safety) is a currently active area of research.

27 Once the elevator is actually built so as to satisfy this specification, these operating conditions may perhaps be regarded as constraints on this particular elevator.

B. Embeddedness and failure

4.6 Stances towards system failure

In this section we shall examine some problems associated with "system failure". Loosely speaking, this term refers to the failure of the system to satisfy some "specification". More precise interpretations will depend on the particular perspective taken towards the notion of "specification". We earlier introduced three such perspectives, or "stances": that of a designer, that of an external observer, and that of the system itself. Failure of the system to meet the observer's expectations will not concern us here; we shall regard this as "model failure" rather than "system failure". Thus, we shall reserve the latter term for use in connection with the perspective of the designer or that of the system. Our discussion will be organized around two dichotomies which we introduced in Section 2.5: simulated vs. embedded systems, and artificial vs. natural systems (under which heading we include, more generally, adaptive autonomous systems).We shall discuss artificial systems from the perspective of the designer, and natural systems from the perspective of the system (together with some ancillary remarks concerning evolutionary "design").

4.7 The problem of failure for simulated systems

a) Simulating constraints

An embedded system, whether or not it fulfills the intentions of its designer or satisfies its design specifications, cannot fail to satisfy the constraints inherent to embedding, such as physical or chemical laws. Now, simulating a system on a digital computer is quite a different matter, because everything becomes a "game", where physical laws can be suspended or altered, and must be "put in by hand" as part of the simulation process. In particular, the system's environment must be simulated, as must the associated constraints. This requires that the relevant relations regarding the environment and the constraints, as well as the specifications, must be assumed to be "known" *and expressible in a format compatible with computation*..[28] This is, in effect, an attempt to use formal implication, algorithmically implemented, as a surrogate for causal, or material, implication. The causal and temporal relations among the physical events taking place inside the computer or the computer monitor bear no relation to those associated with the actual system embedded in its environment. Surely, these distinctions are commonplaces, but they are well worth bearing in mind, in particular in connection with the question of " system failure".[29]

---

28 This latter condition on expressibility is, in fact, quite restrictive; it is, for example, far from clear that "vague specifications" (cf. Section 4.4b)) can be so expressed.

29 We are aware that the same comments regarding substitution of formal for material implication hold for all traditional modeling in the style of mathematical physics, with no harm done thereby. Nor are we by any means claiming that simulation is of no value to the designer or analyst. Indeed, the expectation would be that an iterated process of "model testing and refinement", including under this heading the formal representation of the environment and of the constraints, would be part and parcel of the simulation enterprise.

b) Failure due to logical inconsistency

What we have in mind here is the inability of the system to take any "action"[30] without violating some specification or other (or some simulated constraint). In effect, the system has no "legal moves" available to it. Here "doing nothing" counts as an action, the "null action", so that "no legal moves" means that even the null action is disallowed. Thus we have a genuine logical inconsistency.

(i) Simulated systems

It is surely possible for a set of formally expressible specifications and constraints to be inconsistent, both in principle and in fact. Indeed, due to problems of "constraint propagation", one would expect it to be difficult to write down a set of such specifications with any assurance as to their mutual consistency. Thus, a formal system characterized by such a set of specifications can very easily enter a state from which no "legal move" is possible. We illustrate this by taking as our system a universal Turing machine, with an "action" corresponding to a move on the machine's tape, which constitutes the machine's environment. Let q be the state of the machine, and x the symbol it reads on the tape. Then the pair (q , x) is a "blocking configuration", i.e., a configuration from which the machine cannot take any action, if M(q,x) (the set of allowed moves of the machine reading x while its state is q) is the empty set. A standard result in the subject (essentially dual to the corresponding result for the "halting problem") states that it is an undecidable question (for the Turing machine builder) whether or not the machine enters a blocking configuration. That is, there is no "effective procedure" for determining whether interaction with given "environments" (i.e., tapes) will lead to blocking configurations.

(ii) Embedded systems

As we have argued in Section 2.14, "inconsistency" cannot occur for an embedded system, in particular a natural system.[31] The behavior of the system may be incoherent, but the system will do something, whether this includes actively modifying or simply violating some "specification". It may, in fact, wind up violating the "fundamental specification", which could include undergoing a physical breakdown. In particular, the simulated system will no longer be a valid model of the embedded system. Perhaps the closest "embedded" analogues of such logical inconsistency take the form of conflict or indecision, for example: (i)The problem of "deadlock" for an actually running computer program, leading to a system "crash"; (ii) Two conflicting instructions (say, "turn right", "turn left") sent "simultaneously" to a robot motor apparatus. Here anything can happen (most likely, one command being received slightly prior to the other), but in any case something will happen, even if the "wrong" thing. (iii) A "simple" organism may need to

---

30 There is a wide variety of interpretations that can be given here to the term "action". For example, we may wish to view the system as a discrete-time dynamical system, with an action at time t corresponding to any of the admissible state transitions of the system as a whole. Similarly, an action may correspond to a transition from one "behavior" to another. A rather different type of interpretation may focus attention on some particular system "component" (or "module") which may engage in a range of activities, of which we may wish to single out some class, e.g., "emitting messages", for consideration as "actions".

31 In fact, we argue that it is a category error to speak of "consistency" or "inconsistency" in connection with embedded systems as such.

"decide" between two attractive stimuli, and may indeed spend a certain amount of time "dithering". Almost surely, one of the two stimuli will eventually exert the stronger influence. But in any case the organism will always be doing something, even if it means disregarding both stimuli.

4.8 Inadequacy of algorithmic models

The foregoing remarks add to our misgivings regarding the adequacy of algorithmic models for autonomous systems:

a) Embedding vs. simulation

Our first point is a comparatively minor one, namely that attempts to simulate an embedded system, while they may reveal the presence of inherent conflicts among specifications, will probably not reveal how these conflicts will actually be resolved "in the world".

b) Non-algorithmic modes of operation

Our second point is more serious, and concerns whether a "robot" consisting of a set of physically embedded sensors and effectors interfacing (with one another and) with a controller that manipulates abstract symbols according to a finite set of rules could potentially function as an autonomous system.[32] By the Church-Turing thesis we may, without loss of generality, assume that the controller is a universal Turing machine. The remarks above suggest that the Turing machine by itself could enter blocking configurations. It then becomes, at a minimum, a question of whether the sensors and effectors can somehow circumvent this problem. We have nothing remotely like a "proof" to the effect that this is impossible; however, simple arguments suggest that there are significant obstacles to getting around the problem.[33]

---

32 We do not claim that the difficulty lies with "symbol manipulation" per se, but with the exclusively algorithmic, rule-based, manipulation of abstract symbols. In particular, this framework is purely "syntactic", with no provision for "semantics". If we were not concerned with "autonomy" this might be less of a problem, since the designer could perhaps assume the "semantic burden", and continually intervene to maintain the necessary "semantic/syntactic" harmony for the system. However, in this setting of autonomous systems it is necessary to insist that "semantics" must be internally emergent. We should make a couple of points of clarification. First, we are not seeking here to make a contrast between the "physical symbol" approach of traditional AI and "connectionist" approaches, which are likewise algorithmic in character. We do not feel that the latter approaches address the internalist semantics issue any more effectively than do the former. Second, the above remarks do not bear directly on the "autonomy of syntax" controversies of linguistics, which actually center on a rather different set of questions (which we shall not elaborate upon here).

In Chapter 7 we introduce a non-rule-based schema for the manipulation of physically-embedded "symbols", as part of a tentative approach intended, in particular, to take account of the problem of internalist semantics.

33 We shall consider two potential routes around the problem, and then show where they appear to encounter difficulty. The first route is for the sensors and effectors to have limited interaction with the Turing machine, so that its "environment" is very simple, and hence unlikely to lead to a blocking configuration. However, this has the consequence that the sensors and effectors are left largely on their own, and must thus function fairly much in a purely reactive fashion, linked fairly directly to the physical constraints of the embedding. The second route is for the sensors/effectors to "advance the environmental tape" themselves if the Turing machine is blocked at the current position. However, this will either require: (i) that the Turing machine be able to "recognize" that it is blocked, and then signal accordingly to the

It may be appropriate to reiterate here what we regard as one of the most telling objections to the view that algorithmic modes of operation are adequate for autonomous systems: namely, our intuition that "vague specifications" (cf. Section 4.4b)) can most likely not even be expressed in forms sufficiently sharp to serve as inputs to an algorithm.

c) "Real" planning vs. "classical" planning

The same objections may be raised against "classical planning" in the AI sense, i.e., the (algorithmic) construction of "plans" by computer programs, based on symbolic representations both of "goals" and of the "world". In a sense, such "classical planning" bears less resemblance to "planning" as done by humans than it does to "automated theorem proving", in a sequent calculus style.[34]

One solution, favored by some workers in AI, is to seek to bypass "planning" altogether, and emphasize "reactive" system architectures. While we agree on the importance of embeddedness and situated action, we feel that this in no way militates against "planning" of a non-algorithmic character. In the same vein, we feel that a correspondingly non-algorithmic mode of "symbolization", enabling the system to perform and evaluate "symbolic" actions, can carry significant adaptive value for the system. These considerations will be emphasized in Chapter 7.

4.9 The problem of failure for artificial systems

a) Failure to meet specifications

As we have seen, an embedded system (natural or artificial) cannot enter a state of logical inconsistency, nor can it fail to satisfy its constraints. However, it can fail to fulfill its specifications. The cardinal system failure (for any system, only more so for an autonomous system) is violation of the "fundamental specification", i.e., to maintain itself as a system (cf. Section 4.4a). Other examples are: (i) Failure of a "part" to function as expected (say, due to poor design, or to improper manufacture, or to normal wear and tear, or to abnormal operating conditions, etc.). Indeed, an artificial system is conceived and built to fulfill certain specifications with a delimited class of operating environments. But the state of the environment may evolve outside this class, so that the system is no longer adapted to it. For example, the temperature of the environment can rise to a point

---

sensors/effectors, or(ii) that the sensors/effectors be able somehow to monitor on their own that the Turing machine is blocked. The first alternative seems impossible to realize, due to the undecidability of "blocking. The second alternative very likely faces the same difficulty.

34 We do not have the same objections to the idea of algorithmic "checking" of plans constructed non-algorithmically (to the extent that the resulting plan can be expressed sufficiently formally), or to the incorporation of algorithmic components into an overall non-algorithmic process. Two analogies may be helpful here. First, the difference between the mode of "planning" by a human chess master and that of a chess-playing computer program. (Here it isn't a matter of winning or losing, but "how you play the game"). Second, the difference between theorem discovery and theorem proving, or even the difference between the modes of theorem proving used by a (human) mathematician and those used by an "automated theorem prover". We do not feel that "intuition" should be placed beyond the pale simply because it cannot be accommodated within an algorithm.

As we remarked in Section 2.5c), there is an ongoing controversy in the field of AI regarding "planning" vs. "situatedness", so that "planning" has fallen somewhat out of favor. As noted in a recent talk by a proponent of planning (supplemented by "reinforcement learning"), the concerns raised within the AI field about classical planning include: computational tractability; closed-world assumptions; inability to respond rapidly to unexpected events; inability to handle stochastic or changing worlds and world models.

where the system is destroyed, or the system may run out of power, because the environment does not provide an energy supply. (ii) Inability of the system to act without thereby violating some specification or other, due to incompatibilities among the specifications, given the underlying constraints. (iii) A somewhat more subtle, but hardly uncommon situation, is that the system satisfies all the explicit specifications in its design, yet behaves in the "wrong" way. Here "wrong" is an intentional word which means that the designer or the user of the system realizes that the actions undertaken by the system in a certain context are not the actions expected, or that he judges that the actions do not fulfill certain criteria he had in mind. That is, the explicit specifications are "wrong", and do not realize the implicit specifications or intentions of the designer or the user. Possibly this state of affairs can be altered simply by modifying the explicit specifications. However, just as likely, it will be necessary to rethink the implicit specifications and to resolve conflicts or make choices among different criteria of value (cf. Section 2.14). Indeed this may require an iterative, or cooperative process of refinement, of both the implicit and explicit specifications, not totally unlike the process of implementation of vague specifications (cf. Section 4.4b)).

b) Examples of failure for artificial systems

Rather than staying at an abstract level, we pass to a series of examples taken from Section 4.5:

Example 1 : traditional optimal control problems

Here, we must make a distinction between the mathematical model and the actual system. The role of the mathematical model is to determine an optimal course of actions, or a strategy, given some criteria. This may well run into logical inconsistency; for example, the final boundary condition may not be reachable from the initial boundary condition. As a specific example, one can consider the landing problem, or the moving of an object in given environment). In this case, the mathematical model would predict that no course of action is available (which is a logical inconsistency). But the embedded system would, in any case, do something. In the case of the landing problem, e.g., the system may "do nothing" and undergo free fall, or "it" may try to use its braking systems as much as it can. In any event, the specifications would be violated by the course of action, in this case perhaps resulting in the destruction of the system.

Example 2 : the thermostat

Suppose the system is a sophisticated thermostat which must maintain the temperature between T_1, T_2, and the level of fuel between f_1, f_2 , and must interrupt the heating system while the tank of fuel is being filled (for security reasons). Then it could very well be that, in a rather cold environment, the level of fuel is almost at its lower limit and hence, because it takes a certain time to fill the tank, the temperature goes below the lower limit T_1, so that the system fails to fulfill its specifications. Ordinarily, security considerations should have priority over everything else in case of conflict. In keeping with this, the designer of the system might have included a specification that the security specification should always be obeyed. On the other hand, the user of the system might well not be happy with the result. Still, the embedded thermostat would do

something (namely interrupt the heating system), and no logical inconsistency would arise.

Example 3 : the railway station

In principle, the railway station is designed so that trains pick up passengers on time. This is their primary specification. But there are also security specifications, in particular the specification that no two trains be on the same platform at the same time. Now it could be that the delay of train n°1 induces a delay of train n°2, because the train dispatcher have to obey the security specification; but, in any case, we would not run into a logical inconsistency.

Example 4 : competition among controllers

On certain aircraft there may be a competition between a human controller (i.e., the pilot) and one or more automatic controllers. The system specifications, as built in to the actual hardware, may include the specification that in certain circumstances the automatic controller has priority over the pilot and should take over control of the aircraft, perhaps with the proviso that if the pilot objects (in case of conflicting "sensory" information) then the artificial controller should make the "minimal" action allowed to it. Again, there is no logical inconsistency, and the system will take some course of action. However, it may end with the destruction of the aircraft.[35]

Do to the complexity of modern aircraft there are circumstances where a human pilot cannot "process" incoming information rapidly enough or respond quickly enough to ensure safe operation. Hence, some role for an automatic controller is unavoidable. However, it is far from clear what the proper trade-off should be, nor to what extent the designers should seek to "hard-wire" the trade-off and to what extent allow the pilot to decide when to relinquish control voluntarily. After all, the amount of time required by the pilot to reach and implement decisions may be unacceptably high.[36]

c) Sources of system failure

Returning to our more "abstract" discussion, we want to list a few of the possible sources of "failure" for an embedded artificial system.

1°) Incoherent evaluation criteria

In general, there is no one canonical criterion for evaluating system performance. More commonly, the designer or user (jointly, or individually) have in mind several plausible, yet mutually incompatible, criteria. Thus, the system may fail according to one evaluation, and be successful according to another. As noted at the end of paragraph a) above, such conflicts are not uncommon at the level of "implicit" specifications. They may, however, also make their way into the explicit specifications. In any case, it will be necessary to resolve these conflicts, for example by making choices as to which criterion

---

35 This example is not of purely academic interest. The situation depicted is believed by some to be the source of the Airbus catastrophe of 1989.

36 This example brings into view two interesting types of design trade-off: (i) the trade-off between situated (as well as non-algorithmic) action by the pilot vs. pre-planned algorithmic control by the automatic controller; (ii) the design decision as to how much decision-making should be done in "design time" and how much choice allowed in "somatic time".

to use in any given context. Depending on the plasticity of the system architecture, instead of having to make this decision in advance, once and for all, it may be possible to make it on a case by case basis as the cases arise.

2°) Incoherent specifications

Similar considerations to those above hold for the "resolution of conflicts" among specifications more generally. (Compare Sections 4.4b) and c)).

3°) Inability to anticipate contingencies

If the system must function in a real-world environment rather than in a highly stylized workplace, the designer will not be able to make an exhaustive list of all possible types of situation that may arise, let alone take them all into account in the design . In particular, extreme situations may have been unforeseen or discounted, and the system may not be designed to deal with these situations successfully.[37]

We will also list under this heading failure to anticipate required functionalities or resources in the "internal" environment of the system. We may perhaps phrase this as an inability to determine the "closure" of the set of initial specifications and constraints under the operations of logical and material implication.

4°) Overspecification

The following may perhaps be regarded as a special case of 2°) above. In certain situations of the environment, the system may need to resolve conflicts over resources: for example, access to effectors. Relative to such situations, the system appears to be "overspecified" because, roughly speaking, two subsystems "compete" for simultaneous access to the same effector system, perhaps by sending out "request" signals. Again, for a simulated system this would lead to inconsistency. But for an embedded system, something will happen in any case. For example, one signal may arrive before the other, or the two signals may be unequal in magnitude, or the two signals may cancel each other, so that neither receives access, or the system may "crash", ... Still, in this situation, the system would not be functioning in a satisfactory manner (perhaps acting at random according to the fluctuations in the transmission of signals). In this case, it is necessary to introduce another "control layer" which decides what to do, i.e., which subsystem should have access to the effector .

In other words, the system, given its current structure, cannot fulfill all its specifications in certain situations. (It is, in some sense, unadapted to these situations). This induces conflicts which are, in any case resolved, but in an "unintelligent" way. To bring about coherent behavior it is necessary for the designer (or the system, if it can modify itself) to introduce additional "control" to assist in regulating access to the resources (here the effector resources). This is tantamount to introducing new specifications concerning the control, or revising the logics of the system (cf. Section 3.8d)).[38]

---

37 All systems fail eventually, if for no other reason than normal wear and tear or aging. This surely should not be attributed to a failure to anticipate contingencies. However, faulty manufacture aside, *premature* failure of artificial systems may fall under this heading.

38 Other reasons, besides conflict resolution, for introducing additional control layers include: new functionality, enhanced efficiency, etc...

d) Dealing with the possibility of system failure

If a part breaks down, it may be necessary to replace the part; whether or not this is possible, the system may need to go over to alternate modes of operation. If the system cannot meet all its specifications it will be necessary (though perhaps not possible without introducing other ill effects) to change some specification. The determination of which changes in specification to make, and how to implement them, will need to be made either by the designer (on the "design time" axis; cf. Section 3.3b)), or by the system itself (on the "somatic time" axis).[39]

The system must be able to "deal with" the possibility of system failure. This involves the capacity for anticipating, preventing, monitoring and/or detecting, and recovering from or adjusting to system failure. This goes hand in hand with the capacity to construct, test, modify, and implement specifications at various "levels". We reiterate here the important distinction noted in Section 3.3b): If specification construction is done by the designer, on the design-time axis, then the designer can generate an actual population of candidate specifications for testing either "off-line" or in the environment. If, on the other hand, the specification construction must be done by the system itself, in somatic time, then either: (i) The system must experiment purely by direct interaction with the environment, and suffer the consequences, or: (ii) The system may draw on the "memory" of previous experience to select a "likely" specification for direct testing, or: (iii) The system may generate "plans" and do plan evaluation prior to testing via actual implementation. In the latter case, the system is still subject to "real-time" constraints, since it must arrive at a decision in a timely fashion.[40] These considerations hold for natural as well as for artificial systems.

4.10 The problem of failure for natural systems

Certainly, biological systems can and do "fail". They can fail to maintain themselves as systems, or they can fail in less conclusive fashions. For example, "parts" can suffer damage, or the system's environmental situation can evolve so that the system is no longer adapted to it, either biologically or socially (take a fish out of water, or put a baby in a math class, or install a monkey at the White House or Elysée). However, the problem of failure takes on a different cast for natural systems (or adaptive autonomous systems, more generally) than it does for artificial systems. This is due to the inherently internalist perspectives that must be taken to the term "failure" (cf. Section 2.8).

a) Internalist vs. externalist perspectives

We say that an organism should be adapted to, or maintain a state of relative adaptation to its environment. Still, observing the organism, we may incline to say that it is taking "wrong" actions, even in "reasonable" environments. Of course we are here introducing our own, external, criteria to judge or evaluate the behavior of the organism. For artificial systems, built expressly in order to satisfy the goals of an external "agent",

---

39 Similarly, in the first example the "repair" (and/or the determination of whether and how to go over to other operating modes) may either be done by the designer or by the system itself.

40 We recognize the highly caricatural nature of the preceding discussion. For example, we surely do not wish to suggest a sharp dichotomy between memory and planning.

external criteria of evaluation are totally appropriate. How else to judge "wrongness" than by external criteria, notably those of the designer, or of some user of the system, or of some third external party suffering the consequences of the system's actions? In the setting of natural systems, such external evaluations are by and large of secondary significance. What matters here is what is relevant to the organism itself, as judged by an internal evaluation system. This is not to say that the organism's "point of view" needs to be wholly opaque to an external observer. The observer , in seeking to "model" the organism, may attempt to externally (i.e., external to the organism) represent the organism's "point of view". This may arise in the context of social interaction ("theory of other minds") or, at the other extreme, in a scientific setting, where one may aim to set up "objective" external criteria consonant with the subjective criteria of the organism. One example of this might be to determine what some of "specifications" of the organism "must" be, given the constraints on the organism; for instance, to take a reductionist illustration, what are the relevant metabolic variables, and what is the range of values within which the organism must maintain them? Such "objective" modeling may to a greater or lesser extent be feasible. However, this is largely besides the point. The organism, being autonomous, must take care of itself, without the assistance of the external observer, and hence is obliged to make use of an internal system of evaluation, just as it must rely on its internal physical resources. Moreover, as we have argued in Section 3.3, this system of evaluation must be tied to the fundamental "proto-goal" of maintaining itself as a system, and must be grounded in and supported by the organism's functional and structural architecture.[41]

c) The need for an internal evaluation and reward system

To reiterate, any natural system, in order to be adapted to a class of environments, must have the necessary "tools" and "resources" to bring about and maintain this state of adaptation; this includes the sensory and effector apparatus, control organization, etc... which enables the organism to deal with a typical situation of the environment. Among these tools and resources, we regard as crucial an internal evaluation and reward system, i.e., "value system" (cf. Section 3.3). Indeed, in order to be able to maintain a state of relation adaptation to a changing environment, the organism must not be too specialized (or too sharply specified). In particular, this means that it will be necessary for "choices"

---

41 In saying that the evaluation system is "tied to" the fundamental "proto-goal" we do not mean to suggest that entire "value system", including the "higher-order value system" needs to be involved primarily in supporting basic survival needs. Rather, what we have in mind is: (i) The original provenance of the "value system", in "evolutionary time" is rooted in basic survival needs, including adaptivity. (ii) The primary value system (and the basic circuitry more generally) form the kernel out of which the higher order value system is "bootstrapped". Correspondingly, the functional and structural architecture of the higher value system is not disjoint from, but instead intertwined with, that of the primary value system, even when it functions in "higher order" capacities. Thus, a somatically-grounded reward (or "valence assignment") system may support and be intertwined with the "processes" involved with intuition and aesthetics in "abstract" domains far removed from survival concerns. This is consonant with taking an emergent, internalist perspective to intuition and aesthetics. (In emphasizing this intertwining we are, in a sense, making a similar emphasis to that of Damasio's "somatic marker" hypothesis; cf. Section 2.18). (iii) We are emphasizing the emergent character of the value system and, more particularly the "internalist", or "self-referred", or "self-involved", or "self-centered" character of the emergence (where "self" refers to the system). (iv) A related consideration is that of evolutionary "preadaptation", such as the hypothesized preadaptation of pongid and human motor systems for the evolution of human language; cf. the footnotes to Section 3.7c).

or "decisions" to be made on a somatic time scale (cf. Section 4.4b)) to take "actions" neither "prespecified" on an evolutionary time scale nor directly imposed by the environment.[42] For example, the scarcity of resources, say effector resources, induces conflicts for access to effectors. Since the organism is embedded, such conflicts would, in any case, resolve themselves one way or another. However, the resolution thus reached might well be inauspicious, indeed actively harmful, for the organism. The role of the internal evaluation system is to "bias" decision-making to the advantage of and "for the sake of" the organism itself. This includes maintaining overall "coherence" of decision-making and behavior. In keeping with: (i) our "internalist" conception of the evaluation system as emerging from within and "for the sake of" the organism itself, and (ii)the fact that the evaluation system, like every other facet of the organism's organization, must be supported by the organism's structural/functional architecture, we conjecture that the evaluation system is bootstrapped by and intertwined with a somatically-grounded "reward", or "valence assignment" system; that is, "knowing" what is good is, at some level, tied to "feeling good" (cf. Damasio's "somatic marker" hypothesis, Section 2.18; see also the footnote to paragraph b) above).[43]

d) Symbolization

Metaphorically speaking, our view is that the evaluation system "announces" to the organism, in a symbolic way, that it could be induced to lose its state of relative adaptation with the environment if it were to pursue the kind of action presently being undertaken (or presently being "contemplated") in comparable situations.[44] The key word here is *symbolic* (used here in the sense of "surrogate"): the organism is spared the consequences of experiencing an actual bad situation, by experiencing a symbolically bad situation.[45] As a result, the organism may be led to change a particular decision, or, if necessary, to modify its decision procedures[46] and even its own evaluation procedures (so as to maintain overall coherence of evaluation, decision-making, and behavior). In

---

42 In speaking of "choices" or "decision-making" in this setting, we are, of course, not thinking here in terms of a "rational agents" picture. Indeed, what we are calling decision-making may be carried out by "unintelligent" subsystems, engaging in non-deliberative processes which would be not be regarded as "thinking". In speaking of "actions" we are referring not only to actions on the external environment, but also actions ("deliberate" or not) by the system to modify its own internal organization.

43 We want to emphasize that, in speaking of "somatically-grounded reward", we are *not* seeking to base everything on classical conditioning or reinforcement theory, either at an abstract functional level or at the level of underlying "circuitry".

As a separate point, for the moment we will not attempt to be precise as to whether the valence is "attached" to the consequence of the action or to the action itself, or to both via some type of "association" process. Actually, this way of putting the issue is somewhat simplistic, given such considerations as context-dependence. Moreover, it may be misleading in suggesting an inherently "propositional" evaluation and reward system, albeit one which is instantiated or realized via the "circuitry" of the organism. In fact, given the emergent character of the value system, to speak of "assignment" or "attachment" of valence or value should be regarded purely as a façon de parler.

44 The judging of two actions or of two situations as being "the same" involves, we think, the interaction of the "value system" and the "categorization" processes acting in a "reciprocally recursive" (or "chicken and egg") fashion. This is a non-vicious circularity, one of many arising in the context of autonomous systems (cf. Section 4.10e); Chapter 7).

45 "Symbolic" does not imply "abstract". In particular, the "symbolic" experience may, in conjunction with the reward system, generate an actual unpleasant somatic state.

46 We use "procedure" in an intuitive sense, with no connotation of "effective procedure" or "algorithm".

other words, it may begin to learn both decision and evaluation procedures, as well as new behaviors or behavioral "policies" (e.g., that certain behaviors should be carried out or, at least, attempted in certain contexts). Each time the organism is confronted with a significant variation of the situation of its environment, its current policies may become inappropriate, and the actions it carries out (or anticipates carrying out) in accordance with these policies may be evaluated negatively (and perhaps forestalled) by the evaluation and reward system. The evaluation process will , via its "symbolic" valence assignment, bias the system away from (re-)experiencing bad situations in the new environment, and induce it to learn new procedures and policies if necessary.

To summarize, we can say that a biological system must be able to face rather large variations in its environment, so that it must be a pluri-potential system capable of achieving relative adaptation to a large class of situations. Because of this requirement, it cannot be completely prewired and preprogrammed. What should be prewired are basic intrinsic evaluation processes (perhaps themselves subject to modification in somatic time) which, symbolically, can reinforce or inhibit courses of actions which tend to lead the system out of its state of adaptation, so that the system does not actually experience real maladaptation, but only symbolic maladaptation.

In speaking here of substituting symbolic for real experience we do not wish to go to an extreme. First of all, "learning from direct experience" cannot be altogether circumvented.[47] A "learning period", involving a certain amount of "experimentation" and active exploration, may be needed to adapt to a new situation. Moreover, even in ostensibly familiar situations, "plans" can fail to work out as expected, indeed may even fail to be "executable", despite being tested and evaluated "symbolically". This means that the evaluation system must be capable of accommodating the vagaries of "real-world" as opposed to "classical" planning, and have a correspondingly non-algorithmic character (cf. Section 4.8c)).[48] In particular, the evaluation system must be monitoring and revising itself as necessary so that, over time, there is not an excessive degree of mismatch between the symbolic testing and evaluation of plans on the one hand and the evaluation of actual outcomes of plans as actually put into effect, on the other hand. (Of course, the burden of change here is on the "symbolic" portion of the evaluation system).

e) Learning and memory from an internalist perspective

We feel that in the context of natural systems (or, more generally, adaptive autonomous systems) an internalist perspective must be taken also to learning and memory. The issue goes deeper than simply the distinction between supervised and unsupervised learning typically made in the context of "machine learning". Not only does the organism (or autonomous system) need to learn without supervision, but what is

---

47 Indeed, at least for humans, exposure to novelty is an expected and desired part of living.

48 We should perhaps clarify two points regarding our use of the terms "learning" , "planning", and (at least implicitly) "memory" in the above discussion. First, while we regard these terms as convenient intuitive labels, we are by no means claiming that they each correspond to a single undifferentiated class of domain-independent "cognitive processes" or structural/functional mechanisms (cf. Section 7.9 b)). In particular, we do not wish to propose a single "operational" definition or characterization of "learning". Second, we are not making extravagant claims as to the learning or memory capacities of lower organisms (although even *C. elegans* exhibits some forms of both associative and non-associative learning; cf. the references in the Working Paper cited in Chapter 1). We expect lower organisms, by and large, to be limited to a *primary* value system, capable of at most limited modification in somatic time.

learned and remembered must be relevant to the organism, as "judged" by the organism's value system, both primary and higher-order.[49] Thus, not only does the value system "determine" what is worthwhile to "do", but also what is worthwhile to learn and to remember. This requires some prioritization, since the capacities for learning and memory are limited resources. This is a reciprocal bootstrapping process: On the one hand, the current state of the value system (including "goals" and "plans") "determines", given the current situation of the organism in its environment (including the organism's internal state), what is currently most relevant for the organism to learn or remember. On the other hand, the capacities for learning and memory, as well as the actual "substance" of what has been learned or remembered thus far, are themselves determinants and "constituents" of the current state of the value system.[50]

Such an internalist perspective to learning and memory differs from the extrinsic evaluation procedures used in the context of (non-autonomous) artificial systems, whether the learning is done by the designer (in "design time"), or by the system in somatic time (assuming the system has been provided with "learning algorithms").

4.11 Concluding remarks

a) Artificial vs. natural systems: specifications and evaluation

For artificial systems, the basic concept is that of "specification". The notion of "specification" summarizes both the purposes of the system and the basis for evaluation of the actions of the system, but it is linked to external criteria of evaluation, either those of the designer or of the user. By contrast, for natural systems (and, more generally, adaptive autonomous systems) the notion of "specification" is somewhat contrived, except in an analogical sense. A natural system has its own intrinsic evaluation processes,

---

49 In stressing the unsupervised nature of learning, we are to an extent disregarding those contexts involving parental or cultural transmission of "knowledge". However, even in such contexts it may, in many instances be more appropriate to regard this less as "supervision" in the sense of machine or "neural network" learning than as: (i) exposing the individual to certain experiences which form the potential substrate for the acquisition of "knowledge" or "skills", and (ii) persuading the individual of their "relevance". This may result in the engagement of active, or at least passively receptive, learning processes ("conscious" or "unconscious")within the individual. (These learning processes draw upon the individual's "background knowledge" and "background capacities". A particularly strong form of this assertion, in the context of language acquisition, is the "poverty-of-stimulus" argument of Chomsky; cf. the reference in the footnote to Section 3.7c).) Here "persuasion" may be effected on an affective or "emotional" basis as well as, or instead of, on an "intellectual" basis. Even the "teacher's" providing positive or negative feedback in response to the individual's efforts may have a "persuasive" as well as a "supervisory" role. (For a discussion of the role of affective processes in the context of "language acquisition" see John L. Locke, "The child's path to spoken language", Harvard U. Press, 1993). Moreover, it may be more correct (though perhaps a bit repellent) to speak in terms of persuading the individual's "value system" rather than persuading the individual regarded as a "rational agent". For example, the primary value system of an organism might be "genetically primed" to respond to certain environmental "releasing mechanisms".

The need to regard memory from the standpoint of its relevance to the organism has been stressed by the cognitive scientist Endel Tuving: "A biological memory system differs from a mere physical information storage device by virtue of the system's inherent capability of using the information in the service of its own survival". (See, M.S. Gazzaniga, "Interview with Endel Tuving", J. Cog. Neurosci., Vol. 3. , Number 1, Winter 1991, pp. 89-94).

50 It would not be surprising if, in many instances, learning what is worthwhile to do and learning how to do it were closely linked. In speaking here of "capacities" we do not wish to suggest that what is at issue is simply "storage" capacity, as if one were dealing with a computer information-storage device.

linked to the fundamental proto-goal of maintaining the system as a system. These evaluation processes serve at the same time both to "define" and to maintain "coherence" of system organization and behavior. One aspect of this is the role of the "value system" in substituting, where possible, "symbolic" for actual system failure and, in particular, sparing the system the necessity of experiencing genuinely dangerous situations which could violate the proto-goal.

b) Maintaining coherence: dynamic vs. static "equilibria"

A key point that we have sought to emphasize throughout this chapter is that "coherence": of behavior, of "specifications", of "evaluation", of "decision-making",... is a "moving target", and that "maintaining coherence" involves a continually ongoing process of monitoring and revision at all levels. In particular, both the system and its "specifications" are emergent, and dynamically evolving. In some sense, the system continually redefines itself while, at the same time, maintaining its "dynamic" if not its "static" identity.[51] Our remarks here, while directed primarily at adaptive autonomous systems, may perhaps serve also as a useful vantage point in reconsidering the process of design of artificial systems more generally.

---

51 What we are saying here is surely in some sense obvious, and certainly would be no surprise to Heraclitus. However, what we are really aiming at, albeit very gropingly, is a scientific/technical framework which can accommodate, without forcing or artificiality, this type of inherently internalist perspective.

5. Some intentional notions for adaptive autonomous systems: behaviors, goals, and values

In this chapter we continue our analysis of basic intentional and phenomenological concepts entering into the description of both natural and artificial systems. We shall start with an analysis of the concept of "behavior". Although it is intuitively very attractive, we shall see that in the setting of genuinely adaptive systems, whether artificial or natural, this concept does not appear to lend itself to formalization of traditional type. This suggests that we shall either need to develop richer formalisms or, alternatively, will be obliged to relegate "behaviors" to epiphenomenal status in the mathematical framework we are seeking to construct. We shall follow up with an elaboration of our treatment in earlier chapters of the notion of "value system", which we regard as critical for a theory of adaptive autonomous systems. We shall, in particular, explore some preliminary steps towards a formalization of the concepts of "goals" and "values".

A. Actions and behaviors

5.1 Actions and elementary actions

a) Representing the system from multiple (externalist) perspectives

Particularly when the system in question is an organism, we may wish to examine the system (consecutively or concurrently) from multiple perspectives or viewpoints, representing the system's various facets, components, and organizational "levels". For certain of these perspectives it may be feasible, and desirable, to express these representations formally, in the form of mathematical or computer models.[1] We feel that these various viewpoints are inherently heterogeneous; i.e., it need not be possible, even in principle, to pass to a single, "lowest level", most detailed representation, whose "primitives" form building blocks for the entities in all the other representations. In this sense, among others, our approach is not reductionist. Thus, while we shall informally speak of "levels" of description, we do not intend to thereby suggest a "hierarchy" of decscriptions.

b) Actions

It will sometimes be useful to speak of an "action" by the system, either on itself or on its environment. For example, we made use of the notion of "action" in Section 4.7b), in conjunction with simulated systems. Typically, we shall find it convenient to think of an action as being decomposable, or resolvable into "elementary" actions. We regard neither notion as in any sense intrinsic to the system, but rather, as being very much "model dependent", i.e., depending very much on the viewpoint taken, or "level of description" in the above sense. For example, in some models "actions" occur instantaneously, i.e., without reference to duration, whereas in others duration is an

---

1 Cf. the integrative modeling framework described in the *C. elegans* Working Paper referenced in Chapter 1. For any given "informally" expressed representation there need not be an intrinsic, or canonical, formal representation to associate with it.

important factor. As an illustration, we consider a robot or an animal moving around in its environment. At one rather coarse-grained level of description, we could say that a "unit" move forward is an elementary action and also a small rotation of the body of the robot or of the animal is an elementary action. But at a more detailed level of description it is clear that this "elementary" action is extremely complicated because it involves flexion or extension of a certain number of muscles, and then, of fibres of muscles and we could go on, down to the biochemical level. In particular, the concept of "elementary action" depends of what we choose to define as "elementary" in this representation of functional architecture of the system.

5.2 The intentional notion of a behavior

a) Behaviors as intentional actions

Greatly oversimplifying, one may say that there are two divergent approaches to experimental biology. The first emphasizes experimentation in the laboratory under strictly controlled conditions, using experimental paradigms designed with strict replicability and "quantifiabilty" in mind. This approach, emulating that of experimental physics, has obvious strengths. However, from the standpoint of the integrative organization of an animal's behavior, it has serious drawbacks. These center on the fact that it is the experimenter, rather than the animal, who decides what the animal can do and how the animal is allowed to do it. In reaction against these experimentally imposed constraints on the animal, the "animal ethologists" emphasized the importance of studying the animal acting freely in its natural habitat, and of seeking to understand the animal's behaviors from the standpoint of their relevance to the animal itself and its needs.[2] Perhaps under the influence of this ethological tradition (and perhaps simply because it is the intuitive thing to do), one tends to describe the behavior of an animal in goal-directed terms: food-seeking, avoiding-nuisance, escaping, nest-building, wandering, etc...

It is clear that this characterization of the behavior is an external observer's description and interpretation of the actions of the animal in certain circumstances. Based on observations of a course of actions and their results, the observer "infers" the underlying "motivations", such as hunger, fear, ... Thus, "causality" in the context of behavior is linked to "motivation".[3] This illustrates the strong intentional connotations of the term "behavior". In some sense, the observer lends to the animal his own "intentions" and "motivations", in effect constructing a putative "internalist" model of the animal.

---

2 See, for example, N. Tinbergen, "The study of instinct", Oxford U. Press, 1951. "Acting freely" is not the same as acting primarily "volitionally". The ethologists, in their concern with the animal's "motivational" states, placed great emphasis on the importance of evolutionarily or genetically "hard-wired", albeit highly complex, instinctive behaviors or "reactions", controlled via "innate releasing mechanisms" triggered by environmental "sign stimuli" (perhaps themselves of a "gestalt" character). (An interesting, though by now quite dated, proposal for a hierarchical integrative architecture for instinctive behavior, based on these "innate releasing mechanisms" is put forward by Tinbergen in his book. While acknowledging its own provisional character, he is at pains to contrast the comparative realism of his own approach, which emphasizes the complex physiological mechanisms underlying instinctive behavior, with earlier attempts at synthesis, such as Pavlov's reflex theory and Loeb's tropism theory.)

3 This is not to say that one disregards physiological, hormonal, etc. correlates or "causes" of the various motivational states.

Correspondingly, "behavior" is attributed to artificial systems only in an analogical sense, because one feels odd in attributing "states of mind" to a system that one knows to be artificial. In this respect the notion of "behavior" is complementary to that of "specification": one speaks of "specifications" in connection with artificial systems, and only in an analogical sense for natural systems.

b) "Classification" of behaviors

It does not appear feasible to seek to "classify" the behaviors of an organism on any "sharp", operational basis. For example, in some settings the "natural" criteria might be centered on "goals", whereas in other cases it is more natural to distinguish a behavior in terms of which "apparatus" is used to carry it out, and the mode in which it is used. Nor, in general, is there is such a thing as a "pure" behavior, distinguishable as such independent of context. As soon as the environment becomes complex and rich, it is rather difficult for an external observer (even if he can observe the internal workings of the organism's functional architecture) to divide up sequences of actions into "behaviors". In particular, even allowing for "overlaps", it is far from clear where one behavior ends and another begins For example let us assume that an animal sees or smells a piece of food and goes towards it. But, on its (direct) way, it meets a repulsive stimulus such as a zone of high temperature. Then, it will deviate from its initial direct trajectory, go around the repulsive stimulus zone and, when it is on the other side, continue on a direct path towards the food. Here we might seek to describe a certain sequence of actions based upon relative salience of various sensory inputs and, intuitively, we could say that the animal first exhibits a food-seeking behavior. If we know that the animal reaches a repulsive stimulus zone, we might then say that the animal begins an avoidance behavior and, finally, we might say that the animal resumes food-seeking behavior. But what if we do not recognize that the animal has encountered a repulsive stimulus zone? After all, we may not have a complete list of repulsive stimuli (even assuming that it made any sense to speak in terms of such a list). In any case, how shall we decompose this sequence of actions neatly in two distincts behaviors ?[4] It is not clear that, placed in exactly the same circumstances a second time, the animal would follow the same course of actions, because it is impossible to place the animal a second time in precisely the same circumstances, due to the fact that it is quite difficult for the observer to control exactly its internal "physiological" (and "motivational") state; above all, the fact that the animal has already undergone a first experience of these circumstances, will probably have modified its future behavior in the same circumstances.[5]

We do not deny that "behaviors" have some reality, nor that (say, in invertebrates) there are stereotyped "behaviors". However, these "behaviors" do not have extensional

---

4 Cf. also the reference to Mpitsos and Soinila given in the footnotes to Section 7.7b)

5 Certainly, with an increase in the amount of experimentation and analysis done by the observer there is reason to expect improvement in the verisimilitude of the behavioral models constructed. All the more so, since we do not insist that the observer's data be limited only to "external" behavior vs. "internal" functional organization. In particular, the models used need not be as crude as that mentioned in the text. For example, "behaviors" might be correlated with basins of attraction, or transitions between basin of attraction, of parametrized families of dynamical systems. However, such characterizations of "behaviors" are highly model-dependent, and in no sense canonical. They do not lessen our concerns regarding the difficulties of principle, rather than of practice, encountered by the observer/modeler in attempting to use "behavior" as an operational concept.

descriptions, nor can they be broken up into "equivalence classes" in the formal mathematical sense. We expect that the "natural" way to approach the notion of "behavior" is from the genuinely "internalist" perspective of how the animal "generates" or "specifies" its own behaviors, and this leads back to the notion of "vague specification" (cf. Sections 4.4a) andb), and the references there to other sections; cf. also Section 3.7c)).

c) Modification of behaviors via learning

Another fact is that a natural system learns. We illustrate with a couple of simple examples of associative learning: Consider an animal initially equipped to detect sensory stimuli 1 and 2 (say, for example, two odors) separately. As the animal evolves (in developmental or somatic time), it may turn out that only the association 1+2 is relevant. In the same manner, it may be that at the beginning of its life 1 is attractive and 2 is repulsive but that, due to certain regularities in the environment, 1+2 are often associated. Then it may turn out that 1 becomes the signal of 2, and so is by itself sufficient to generate an escape behavior.

5.3 Behaviors vs. classes of sequences of actions

One might be tempted to define a behavior as a class of sequences of actions. But there are two major criticisms of such a definition. The first criticism was discussed in the preceding section : In the case of the animal going towards a piece of food while avoiding a repulsive stimulus, which subsequences of actions should be attributed to one behavior (food-seeking) vs. the other (repulsive stimulus avoidance)? Now, we could say that the whole sequence of actions is characteristic of the complex behavior "food- seeking and repulsive stimulus avoidance", but then, for purposes of analysis, how could we represent this more complex behavior as the "composition" of the two simpler behaviors ? The second criticism comes from the following example: Suppose the animal is wandering and, by chance, finds a piece of food and eats it, although the animal is not particularly hungry. How are we to designate this sequence of actions? It is not food-seeking behavior. Still, based purely on our observations of the animal performing this sequence of actions, we may think that the sequence is a sample of food-seeking behavior, because we may attribute the fact that the animal has eaten the food to a possible hunger at the onset of the sequence of actions. This illustrates how a sequence of actions can constitute a sample of a particular "behavior", or not, depending on the context, the internal metabolism, the history of the system etc... We see, once again, that the notion of "behavior" does not lend itself to "sharp" specification, and is not useful as an operational concept.

5.4 Behaviors as emergent [6]

a) Felicitous sequences of actions [7]

As we have discussed in previous chapters (and as we shall elaborate upon in Section 5B), we may expect an organism to have embodied within it an internal evaluation and reward system (or "value system") which serves to maintain the overall coherence of the organism's behavior. Against this backdrop we may speak of the "goals" and "motivations" of the organism, with the understanding that these may "belong" to the value system rather than to the organism regarded as a volitional agent. As a consequence of its goals and motivations, the organism performs certain sequences of actions which we typically qualify as "behaviors". We have emphasized in Section 2.14 that in the setting of embedded systems it is more appropriate to think in terms of "coherence" rather than "consistency". Similarly, it more appropriate to speak of the "felicity" rather than the "well-formedness" a given sequence of actions by the organism, the determination as to "felicity" being made by the value system itself.

Precisely how the value system carries out such determinations will of course depend on the embodiment of the value system within the overall functional architecture of the organism, and this will in turn depend on the particular class of organism under consideration. We will defer to a subsequent paper the construction of "model circuitry" realizing aspects of the schematic functional architecture which we begin to examine in the present paper (see, in particular, Chapter 7). However, to make the present discussion somewhat more concrete, and to emphasize the embodied nature of the value system, we shall provisionally give a metaphorical "neuronal" characterization of the operation of the primary value system (or more specifically, the "component" concerned with homeostatic regulation): The "goal" is to maintain the "firing" of certain neuronal populations within a "baseline" ensemble of firing patterns.[8] In particular, if the firing pattern deviates from this baseline, it is necessary for the organism to act so as to restore a baseline pattern. From this vantage point, a given sequence of actions is felicitous or not depending on whether it results in the satisfaction of this goal.

b) Behavioral architectures of artificial systems

The preceding discussion reinforces the arguments of Section 2.10 as to the inherently emergent character of the "behaviors" of an adaptive autonomous system. This implies now the following consequence. If we want to design an artificial system which is truly adaptive and autonomous, then we cannot impose upon the system architecture sharply prespecified behaviors, nor a fortiori a fixed repertoire of behaviors ordered according to a fixed priority hierarchy or coordinated via fixed "arbitration schemes" (cf. the discussion of the "subsumption architecture" in Sections 2.10 and 7.11a)). The same remarks would hold for attempts to transpose such "behavior-based" approaches to the description of natural systems. Within the functional architecture of an organism we do not expect to find the counterparts of a hard-wired "wandering module" or "edge-

---

6 Cf. Section 2.10.

7 We borrow this terminology from John Austin (cf. the reference given in the footnote to Section 2.14).

8 As noted earlier we are using the term "neuronal firing pattern" in a somewhat metaphorical sense here. (See also the first footnote to Section 5.4c)).

following module" or "shelter-seeking module", specialized to produce "wandering behavior", "edge-following behavior" "shelter-seeking behavior", etc... as building blocks of the overall "behavioral architecture" of the organism.[9] It may be that for simple "robots" or "artificial insect models", one can design and implement a behavioral architecture in this way, and that it will even work successfully according to some (external) evaluation scheme. But this does not mean that the system will be adaptive and autonomous: If the designer prewires once and for all the arbitration scheme and/or priorities among behaviors, the system will ipso facto be unable to learn new priorities, let alone new behaviors, and will be unable to adapt itself to a new class of environments. Alternatively, still remaining within this general design framework, the designer may allow the system the possibility that the hierarchy of priorities is not fixed. But then, the changing of priorities would be behaviors in their own right, for which the designer would need to specify priorities, just as for the original hierarchy; e.g., when does a change of priority have priority over maintaining the original priorities ? [10] Moreover, if changes of priority are permitted, certain prewired behaviors may become obsolete and be abandoned as the system evolves (in somatic time). The corresponding "hardware" modules specialized for the production of these particular behaviors would then become useless, yet remain intact and unco-optable for other purposes. This is very different from the picture one has of an actual nervous system in operation.

c)Basic prewired circuitry

As regards "emergence" (on a developmental or somatic time scale) for actual organisms, it is not that we deny an important role for "prewiring", as should be clear from our discussion in earlier chapters; see, for example, our discussion in Section 3.3 of the "basic circuitry" already present as part of the "circuitry" of the newborn animal. However we think it unlikely that this "prewiring" is done on a "behavior" by "behavior" basis. Speaking loosely, it might be reasonable to view some "behaviors" as more prewired than others; which behaviors these are is largely an empirical question, depending on the specific species of organism in question. Also, we wish to distinguish between "prewired" and "hard-wired", the latter term suggesting "sharp specificity".[11]

---

9 We recognize (as we shall elaborate in Chapter 6) that there are various distinct interpretations that can be given to the term "module" (and to the associated notion of "modularity"). In particular, we do not claim that the modules in a behavior-based decomposition would need to correspond to the modules in a function-based decomposition. The point we are suggesting is that a behavior-based modular architecture of the type described in the text could not serve as a reasonable approximation, even at the strictly behavioral level, to an organism. Our argument here is based largely on the issue of "emergence", and leaves aside the separate set of concerns focusing on the "reactive" vs. "representational" emphasis underlying such behavior-based approaches.

10 The simplest case is when there is a prewired context-dependence in the original hierarchy (see Section 2.10).

11 We should perhaps clarify a couple of points: (i)Even with "vague specifications" the resulting outcome may have a good deal of "precision", with comparatively limited context-related variability. What is at issue is more the "unspecification-like" character of the "specification" process than the degree of variability in the outcome. An example from an embryological context is the differentiation of cells into distinct cell or tissue types. In some organisms such as *C. elegans* this process is highly stereotyped, yet is nonetheless subject to influence by the local context of the individual cells. (ii) We do not intend such terms as "circuitry" or "wiring" to be taken in too literal a sense. To begin with, even granting that we are speaking of an organism with a nervous system, we do not want to make too sharp a segregation between the nervous system and the other structural and functional "subsystems" of the organism. For that matter,

This brings us back to the issue of "vague" vs. "sharp" specifications, including at the level of genetics and development. (See also the discussion of "diffuseness" in the basic value system, Section 3.3d)).

Similarly, "emergence" is compatible with significant "specialization" of structure or function within the resulting architecture. For example, portions of the basic circuitry can be quite specialized for producing certain sequences of actions, or at least are almost "primed" to become so. In particular, those portions linked to basic survival needs are likely to fall within this category. More generally, specialization and "localization" of function is a basic characteristic of nervous system organization for both higher and lower organisms.[12] A notable example is provided by the highly specialized "language areas" of the human brain.

B.Goals and values

5.5 The intentional concepts of goals and values

We emphasized earlier (see Sections 3.3, 4.10, and 4.11) that an adaptive autonomous system (notably, an organism) has its own goals, in contrast to a non-autonomous system, whose "goals" are externally imposed by the designer. Correspondingly, the autonomous system has an internal evaluation and reward system, or "value system", linked to and arising out of the fundamental proto-goal of the system, namely to maintain itself as a system. This means, in particular, preserving its autonomy while maintaining a state of relative adaptation with its environment.

As with the notion of "behavior", such notions as "goals", "evaluation", "reward", and the like have a strong intentional connotation. Again, we do not envision a fixed list of goals or evaluation criteria, nor a fixed priority hierarchy among various goals. Moreover, the system's goals are clearly context- and history-dependent. In particular, goals do not exist in isolation, but within the context of complementary or competing goals, as parts of plans or strategies.[13] Certain goals become obsolete, other goals emerge. As the situation of the organism evolves, what was initially a "high-priority" goal, may quickly become relegated to secondary importance.

In seeking to formalize the notions of "goals" and "values", we face similar difficulties to those associated with formalizing the notion of "behaviors". However, the former notions are more overtly, as opposed to implicitly, internalist in character than the notion of "behaviors". This suggests that they may be more amenable to formalization than the latter. Following up on this intuition we shall, in the next two sections, take some

---

one expects much of the "basic circuitry" (e.g., involved in homeostatic regulation) to be non-neural in character, with the neural portions integrated into the overall "control " organization. Moreover, even in focusing on the nervous system per se, we do not wish to overemphasize exclusively synaptic modes of "signaling" and communication at the expense of more diffuse modes of interaction, including hormonal and peptide signaling. In particular, we want to take account of modulatory action.

12 Although one should not interpret this in too "naive" a fashion.

13 As noted earlier, these "goals" or "plans" may "belong to" the value system rather than to the autonomous system as a "volitional" or "rational" agent. In particular they may, to an extent, constitute part of the evolutionary or genetic endowment of the system (with these terms suitably interpreted in the case of an artificial autonomous system). Similarly, various "choices" may more properly be attributed to the value system than to the autonomous system as a "volitional agent".

steps towards a general scheme for an (internalist) formalization of the notions of "goals" and "values", expressed in terms of the structural/functional architecture of the system. For concreteness, we take the system in question to be an organism, more specifically an organism with a nervous system, so that we can continue to use the "neuronal" imagery of Section 5.4a).

5.6 Primary goals and values

a) The fundamental goal of an autonomous system

As we have observed in Section 3.3, every autonomous system has an implicit, system-referred fundamental goal (or "proto-goal" or "specification"), namely to maintain itself as a system. This requires that the system maintains a state of relative adaptation with the environment. In particular, it requires that the system satisfy the "specifications" associated with the basic constraints discussed in Section 4.2, in particular the homeostatic constraints. As discussed more fully in Section 3.3, the satisfaction of the "fundamental" goal, together with the "specifications" and "goals" concomitant to it (which we shall refer to as the "primary goals"), is supported by the "basic circuitry" of the system. This "circuitry" incorporates, in particular, sensory apparatus, sensitive to features of the external and internal environment, and effector apparatus allowing the system to change its relative situation in the environment, as well as to modify its internal organization. It also provides the substrate for the "primary value system", which acts to "bias the system" to actions tending towards the satisfaction of the above "specifications". (In the present paper we shall more or less take for granted the fact that the basic circuitry is in place and functioning. We shall defer a detailed discussion of its structure and function, including how it deals with competition among various goals, to a subsequent paper in which we shall also experiment with some "model" circuitry.)

b) The primary goals and values

It will be convenient to use a single schematic to represent the various forms of homeostatic regulation problem faced by the system: The set of relevant variables (e.g., certain metabolic variables) may range over a certain domain D in an abstract space (which we do not wish to specify further), and the corresponding "primary goal" of homeostatic regulation is to maintain the metabolic variables inside a certain subdomain D_1 of D. Depending on the particular homeostatic problem under consideration, this subdomain need not be fixed, but may to some extent be context-dependent.

Examples of homeostatic regulatory variables include, in addition to metabolic variables (e.g., the chemicals needed for the energy supply of the system), variables such as the temperature of the system, or variables expressing information about the state of various parts of the "body" (such as kinesthetic information), or chemical or physical substances that the system should avoid, so that their value should be maintained very low (for example poisonous substances or unpleasant substances or dryness...). We are not concerned here with making a comprehensive list of "all" the homeostatic variables relevant to the system, even assuming this were feasible in principle. Aside from depending on the particular class of organism considered, such a list would be highly dependent on the types or "levels" of models used to represent the organism.

As discussed in Section 3.3 there are also other "primary goals" of the system not directly corresponding to homeostatic regulation. These include evolutionarily derived "biases" towards certain patterns of action: for example, avoidance of predators, attraction to various stimuli, ... Some of these other primary goals may have indirect links to goals of homeostatic regulation.

c) Realization of the primary goals and values

In connection with the problem of homeostatic regulation, we shall assume that the prewired basic circuitry contains apparatus to carry out the following functions:(i) "Monitor" the homeostatic variables in question.(ii) "Signal" when these variables are entering a dangerous domain D_2 containing D_1. (iii) Generate corresponding actions tending to return the homeostatic variables to the allowed region D_1. As mentioned in Section 5.4a), we shall present a somewhat metaphorical picture of the associated processes in terms of "neuronal activity". Thus, we shall assume that the above signaling is realized as the deviation of a certain class of neurons (perhaps an isolated neuron or, alternatively, a neuronal population) from a "baseline firing pattern"; this signaling may also involve the release of hormonal or peptide modulatory substances. The primary goal is to effect the return of this firing to a baseline pattern via appropriate interaction with the environment. This, in particular, relies upon the constraints associated with the embedding of the system in its environment; for example, the ingestion of certain substances induces a rise of certain metabolic variables (due to biochemical laws), and this in turn induces a reduction of the deviation from baseline of certain neural firing.[14] The value (or reward or evaluation) of the environmental interaction may itself be realized as the approach towards a baseline pattern of firing of yet another neuronal population; we picture this as a population whose degree of deviation from a baseline pattern is correlated with that of the neuronal population generating the signalling. (This "correlation" results from the overall pattern of connectivity of the basic circuitry).[15]

5.7 A basic "strategy" of adaptive autonomous systems : symbolization[16]

We feel that processes of "symbolization" figure prominently in the functional architecture of adaptive autonomous systems (cf. Sections 2.19 and 4.10d). As we use these terms, both "symbols" and "symbolization" are embedded in and emergent from the structural/functional architecture of the system. In particular, "symbolization" does not correspond to the abstract syntactic manipulation of disembodied, abstract "symbols" of the kind associated to formal languages. Moreover, we wish to give rather broad scope to the term "symbol", so that it in particular might carry some of the non-formalist, non-AI

---

14 As emphasized in Chapter 4, for an embedded system, the constraints will necessarily operate, and the correct interaction with the environment will automatically induce its natural physical or biochemical effect. On the contrary, for a simulated system one has also to simulate the environment and the constraints.

15 Of course the environment may be sufficiently nasty (e.g., not providing adequate food sources) so as to overwhelm the capabilities of the homeostatic apparatus to maintain the relevant variables within the normal range D_1. For example, in the case of metabolic variables the natural tendency (due to metabolic constraints) is for the variables to go down and enter the dangerous zone D_2 , unless the homeostatic apparatus can compensate. (Since the system is embedded something will "happen" in any case).

16 The term "strategy" is not intended to suggest a "volitional agent".

connotations of the intuitive notion of "representation" (or "internal representation").[17] We shall discuss "symbolization" at length in Chapter 7 (where we put an emphasis on the nervous system, and use the terminology "neural symbolization"). However, we wish to give a preliminary discussion in this section and the next of certain of the roles played by symbolization.

a) Surrogate symbols

The general idea is that an adaptive autonomous system tends to develop structures for generating and for manipulating "symbols" of actual experience. Via manipulation of these "symbolic surrogates" the system tries to predict and avoid actual "bad" experiences. In the same manner, the system tries to predict and seek out actual "good" experiences, while avoiding the costs (in terms of energy, time, etc,...) of actual search of the environment.

b) Contrast with the reflex arc

It may be instructive to contrast this use of "symbols" by the system with the purely reactive behavior associated to reflex arcs. In the latter case an actual stimulus, in particular a stimulus of negative value, does indeed generate an instantaneous reaction, but it should be noted that this stimulus was a "surprise" to the system and that the system has had to undergo an actual experience, typically a "bad" one. Similarly, such a purely reactive "strategy" would be a very costly way of giving rise to "good" experiences.

This suggests an interesting slant as to the "selective advantage" conferred on the system by having sensory modalities sensitive to "signals" generated at a distance. Similarly for association, learning, memory, planning and language.[18] These all circumvent the necessity of dealing directly with the environment, as well as permitting a wider range of behavioral choice than would otherwise be possible. They allow the system to build a symbolic environment and "play" with it before committing to specific courses of action in the actual environment.

---

17 In the present paper we say very little about "representations" per se, or about the processes by which representations are formed, manipulated, and modified. This is not because we consider "representations" as unimportant. Quite the contrary. However, it is not clear to us at the moment precisely what the relation "is" between the notions of "representation" and of "symbol" (or "neural symbol"), although they seem very closely linked. We feel that "symbol", at least as we use the term, is sufficiently broad to incorporate the various senses of "representation", though not vice versa. Thus we have *provisionally* opted to place the "representational" burden on the notion of "symbol", and to emphasize the process of "symbolization" rather than "representation construction". Again, this is not to say that we regard "representation" as a secondary or derived notion, but merely that we regard "symbolization" as a better starting point for our work. We anticipate that in subsequent phases of our work the notion of "representation" will be better elucidated, and will figure in an "independent" capacity. As we stress throughout this paper, "symbolization", in the sense we use the term, is not associated primarily with algorithmic, or computational, processes. Accordingly, the same will be true of "representations" in our sense of the term. Thus, both "symbols" and "representations" take on a different character in our setting than they do in AI.

18 While language surely adds a whole new "dimension" to symbolization, the processes of "symbolization" we have in mind would pertain in one form or another to essentially all organisms, and need not have anything to do with (human) language.

5.8 Higher order goals or values and their symbolization

a) The primary goals and values as a reactive system

We return to Section 5.6 above and the "basic circuitry" supporting the primary goals and values. It appears that in its role of homeostatic regulation, this circuitry functions to a large extent in a "reactive" manner. That is, certain values of the homeostatic variables may actually need to enter a "danger zone" before the "basic circuitry" is induced to generate an appropriate countervailing interaction with the environment. In such a case, the system not only deals directly with the environment, but actually suffers a "bad" experience, which might potentially threaten the survival of the system. However, there are other cases where the "basic circuitry" very likely makes use of symbolization: for example, the mode for avoiding poisonous stimuli presumably does not allow, let alone require, the system to test the stimuli directly. Instead, what might happen is that a certain repulsive stimulus will act as a de facto "signal" of the actual poisonous stimulus.

b) Building up of higher-order goals and values by the symbolic strategy

In the above example of the poisonous stimuli the "symbolization" was "learned" or constructed on an evolutionary time scale. However, such symbolization can also take place on a somatic time scale. The strategy uses "association".[19] After a repeated number of experiences, the experience (either good or bad) becomes associated to another sensory experience which usually prepares for or announces it due to certain constraints of the system and its environment. A priori, this new sensory experience was neither good nor bad in terms of the primary values described in Section 5.6. However, it itself becomes a new goal, and the reinforcement of the neural firing pattern associated to this sensory experience becomes a new value.[20]

So, for example, a repulsive stimulus may be announced by a secondary stimulus (like a smell or a shape) which can be recognized while the system is still at a distance from the original repulsive stimulus, which can thus be avoided without being directly experienced by the system.

---

19 Here we are using "association" as an intuitive, catch-all term, rather than in any technical sense. We shall defer any attempt at a technical treatment to a later paper dealing with explicit "circuitry". However, we do wish to make a few remarks here to avoid possible misunderstanding: (i)We are not claiming that there is a single, universal type of "association", nor is there a single mechanism or process for forming such associations. (ii) In particular, we are not assuming that one should look for general-purpose vs. domain-specific mechanisms (or even principles) for learning, memory, attention, etc... (iii) Nor do we want to claim that all associations are formed (at a psychological level) via some form of "conditioning" (cf. Damasio's "somatic marker hypothesis", referenced in Section 2.18). For that matter, we recognize that even in the context of classical conditioning (cf. C. R. Gallistel, "The organization of learning", MIT Press, 1990) things are complex: Factors of timing between stimuli or between stimuli and action, etc., play a role. Moreover, the "symbol" is not a complete substitute for the original. (iv)We wish to consider associations between stimuli and actions as well as between stimuli. (v) We wish to consider also the learning of relationships as relationships, as contrasted with simply learning instances of the relationship.

20 We are speaking somewhat loosely here in order to emphasize the main point. Certainly, a more careful discussion would need to deal with many factors not touched upon here, e.g., with the effects of "context-dependence" on the firing pattern, with the relation of the neural substrate of the value system to the "sensory areas" supporting the sensory firing pattern, etc... Indeed, as before, we use the term "neural firing pattern" in a somewhat metaphorical sense here.

In the same way, a "good" stimulus can be announced by a secondary stimulus which the system can recognize while it still at a distance from the original stimulus. This makes it unnecessary for the system to start a random search for the good stimulus. Such a search could be very costly in metabolic or other terms and might , in particular, cause the metabolic variables to approach dangerously low values (as a result of the metabolic constraints). After this kind of association is developed, the secondary stimulus becomes a goal in itself and the corresponding neural firing pattern serves as a symbol for the original stimulus. Correspondingly, the reinforcement of this firing pattern becomes a new value. Thus, the system will try out certain new kinds of interaction with the environment, in an attempt to bring about the pattern of neural firing corresponding to this secondary stimulus.

Underlying the above picture of symbolization is the assumption that the structural/functional architecture of the system, using the basic circuitry and primary value system as a kernel (cf. Section 3.3), has the apparatus necessary to form such associations and to generate the corresponding elaborations of the value system. But, if so, this process of symbolization and building of higher-order goals and values has no reason to stop after one iteration.[21] Moreover, as the system evolves (in somatic time), certain new goals will receive "higher valuations" than others, and other goals may have their valuations diminished, depending on the kinds of environments the system experiences. It is important to remember here, as stressed in Section 5.5, that individual "goals" cannot be treated in isolation, but must be regarded as parts of larger, coordinated patterns. Similarly, the "comparison" of two values is best regarded from an internalist perspective, as a complex context- and history-dependent process carried out by the system via the structural/functional apparatus available to it (in particular its internal evaluation system), and not simply a straightforward rank ordering of "scalar" quantities according to some external criterion.[22] Again, this raises the important issue of how "coherence" of goals and values is maintained (cf. Section 4.10).

c) Richness of the environment and of the system

This strategy of "symbolization" and building of higher-order goals and values can work only for a system which evolves in a "rich" environment and whose evolving structural/functional architecture is capable of supporting rich sensory experiences and associations. In particular, the capacity for sensory experience "at a distance" is especially important in connection with symbolization. We conjecture that such a capacity would naturally co-evolve (on an evolutionary time scale) with the structural/functional architecture for forming "internal representations" of the external world.

---

21 We recognize (cf. the reference to K. S. Lashley in the footnotes to Section 3.7c)) that learning temporal chains of actions involves more than simply learning a sequence of successive pairwise associations. Similarly for making A_n into a symbol for A_1 simply via an iterated chain of intermediate symbolizations.

22 To illustrate the inappropriateness of looking for a canonical, context-independent ranking criterion: Is ranking according to "importance" or according to "urgency" of higher priority ?

6. Some basic analytical notions for autonomous system description

In this chapter our emphasis will be on methodology rather than on underlying principles of system organization. That is, we shall be concerned with language for describing system organization rather than with working hypotheses as to actual organizational principles. We shall examine three basic notions: "state", "module", and "control", which we think will play a role in the analysis and description of adaptive autonomous systems. As may be surmised from our earlier chapters (see, e.g., Sections 2.11 and 2.6d)), and as we shall discuss further in the present chapter, we do not think that "traditional" treatments of "state" or "control" are well-suited to the description of adaptive autonomous systems. Thus, we wish to attach non-traditional interpretations to these terms. The notions of "state" (or "internal state") which we have in mind are linked to the "value system". This includes both "global" perspectives, associated with the system as a whole, and "local" perspectives associated with subsystems. We envision a very broad and flexible notion of "module", as a potential means of formally representing various types of informal "subsystem decomposition". In particular, the framework of modules communicating and interacting with other modules will provide a setting in which we discuss various modes of "control". The most interesting of these, in our opinion, involves the use of symbolization.

We wish to stress that the discussion to follow is still at an informal level. We are not attempting here to actually construct a formal descriptive framework, but only to examine "components" for possible incorporation into such a framework. Since we shall be discussing artificial as well as natural systems, the viewpoint taken will often switch between that of description and that of design.

A. States and modules

6.1 States for Turing machines and for abstract models of computation

a) For an abstract, non-embedded system, like a Turing machine, one can define the "state" as a primitive concept without any semantic content. The Turing machine is then characterized by a finite number of states (or internal states), so that when the machine is in a state q and reads a symbol x on the tape, it functions in a specified way: change the state q, then change x, and then move on right or left along the tape. For a Turing machine the "environment" is the tape. One thus has a purely "reactive" system which, "reading" its environment, changes it after a number of moves.[1] We note that this concept of state has been used by Wonham-Ramadge in their theory of discrete-event control, but in a finite automaton framework (cf. Section 2.7).

---

1 As noted in one of the footnotes to Section 2.13, there are several distinct meanings attached to the term "reactive". Here (as elsewhere in this paper) we use the term "reactive" to refer to the absence of "control structures" mediating the system's response to environmental stimuli. Some might argue that, because it has "internal states" mediating the input-output relation, the Turing machine is not reactive.

b) The Turing machine picture is one of the basic abstract model for computation. A variant of this picture is that of a Turing machine with three tapes: one of the tapes contains the input data (a reading tape), another tape is the memory store, and a third tape receives the output. The memory tape is now to be regarded as an internal part of the machine, and the "state" of the machine is now what is usually called a "configuration", namely the previously defined state q together with the word written on the memory tape. This has the advantage that the description is more "realistic", and the disadvantage that the state space is infinite.

c) In the above picture (deterministic) finite automata correspond to those cases where there is no memory tape. Thus, they have no memory beyond the finite-state memory associated to the space of internal states q. For various purposes this is a significant restriction.

6.2 Modules

a) Artificial systems

Modularity is a basic guideline in the design of engineered systems, whether embedded or formal (e.g., computer software).[2] That is, one seeks to decompose the overall system into a set of simpler subsystems or "modules" with clearly delineated functionalities. The various modules may be concerned with distinct types of functions, but this need not be the case. For example, as a form of parallelism, it may be desirable to have multiple copies of essentially the same module, carrying out identical functions, but at different locations, and acting on different "inputs". One basic feature of modular design, emphasized particularly in computer science, is "encapsulation". That is, the individual modules interact with one another and with the external environment only through specified interfaces.

The decomposition of the system into "modules" is surely not unique. For example, starting from a given decomposition one can group several modules together to form a larger module, or look inside a module and decompose it further into smaller functional or architectural units. Reciprocally, the given decomposition may be a particular form of "coarse-graining" (or "projection") of a more refined decomposition, involving not only a larger number of modular units, but also more refined interfaces or channels of communication than the given decomposition. Moreover, the original decomposition need not be "canonical". There may well be multiple viewpoints one may wish to bring, or may need to bring to the system, leading to a distinct modular decompositions, "incommensurable" with one another. This is particularly to be expected for embedded systems, where constraints associated with the substrate supporting the originally envisioned functionalities lead, via "propagation" of effects, to the need for additional functionalities beyond those originally envisioned. Thus, modular decompositions may be only "approximate", as well as viewpoint dependent.[3]

---

2 It is not unreasonable to regard computer software from the standpoint of a formal system even though it must eventually be embedded. For the present purposes what is relevant is the extent to which embedding considerations enter into the design.

3 In the context of the organization of an organism, the need to look beyond "intrinsic" modular decompositions, and to examine the notion of "approximate" modularity, was stressed in Rockland, C.

Similarly, it is important to remember that structural, functional, and behavioral "decompositions" need not correspond to one another in a one-to-one fashion. For example, the supporting substrate of a functional module (within a particular functional decomposition) need not be localized structurally. Reciprocally, a given structural "module" may participate in multiple functions, either simultaneously or at distinct times.[4] Once one takes account of time, one should also consider the potentially dynamic vs. static nature of modular decomposition. Indeed, the modular decomposition may be under the control of the system itself. For example, a "higher order" controller may configure the motor apparatus into quasi-independent modules for various purposes, the particular configuration varying with the purpose. Similarly, the relation between functional "decomposition" and corresponding structural substrate may be dynamic. One good illustration of the dynamic character of modular decomposition is associated with the "communicating, interacting agents" approaches to concurrent computation. If one regards communicating "agents" (or "processes" or "actors") as "modules", then in these approaches "modules" may come into and go out of existence at all times as a result of the interaction among them.

These caveats notwithstanding, we shall for various purposes find it useful to assume that a given decomposition into "modules" has been determined. The modules communicate with each other and with the global ("external") environment of the system as a whole. In discussing this communication, it will be convenient to make use of perspectives localized to individual modules, and to speak of the "proper" environment of a given module, i.e., those modules and/or the global environment with which the module is in direct communication. The role of a module, abstractly, is to receive "inputs" from its proper environment and to communicate "outputs" to its environment, though not necessarily in a reactive manner.[5]

(1989) The nematode as a model complex system: a program of research. Laboratory for Information and Decision Systems, Working Paper (*LIDS-WP-1865*) 1-115 (plus appendices).

The necessity of considering "multiple modularities" even in the context of engineered systems is discussed in Rowley, S.G. and C. Rockland (1991) The design of simulation languages for systems with multiple modularities, *Simulation* , 56:3, 153-163.

4 However, depending on the system being considered and the questions being examined, it may well be appropriate to think in terms of a certain degree of "localization of function", suitably interpreted. This, for example, appears to be the case with the nervous system, provided one interprets "localization" suitably broadly. Moreover, when dealing with embedded systems, it will be advisable to study both functional and structural decompositions coordinately, despite (and, to some extent, because of) the lack of one-to-one correspondence between them.

5 The terminology of "inputs" and "outputs" is convenient, and we shall often make use of it, despite the fact that a more neutral terminology such as "interfaces with its proper environment" might carry fewer unwanted connotations. Among the connotations we wish to avoid (except in special contexts, expressly designated as such) are: (i) The existence of fixed, dedicated input and output "channels". (ii) The assumption that the "channels" are structureless entities serving purely as passive conduits or "buffers". (iii) The assumption of strictly point-to-point communication. (iv) The view of "communication" as a formal vs. embedded process, mediated via abstract "tokens" of homogeneous type. (v) The assumption that all interfacing between modules or between modules and the global environment involves exchange or "extraction" of "information". We want the picture of "communicating modules" to be broad enough to accommodate the various modes of cellular and, in particular, neuronal interaction. Thus, we wish to allow for non-synaptic modes of interaction, including various forms of "modulation".

b) Natural systems

The above caveats hold all the more forcefully in the case of natural systems. Indeed, the emphasis on modularity in connection with engineered systems is associated primarily with "externalist" perspectives, those of the designer or maintainer (or "repairer") of the system. The concern is less with the benefits that modular organization may confer on the system itself than with the resulting ease of description, analysis, fault-diagnosis and repair, etc. This is not to say that a considerable degree of "approximate" modularity (and concomitant "specialization") cannot arise in an emergent fashion, nor that it would not also confer benefits on an autonomous system, as evaluated by its own value system. However, we believe that such approximate modularity is a rather richer and subtler affair than the the type of modularity designed into engineered systems; indeed, we suspect this type of "strict" modularity is responsible for some of the "brittleness" of the latter.[6]

Nevertheless, we think that the introduction of "modular decompositions" can be a valuable means of facilitating our analysis and description of system organization, as long as we are careful to distinguish between "intrinsic" modularity , and modularity which we to an extent "impose" upon the system. In fact, this may be the best route towards elucidating the notion of "approximate modularity". One basic analytical role we envisage for "modular decompositions" is to serve as a backdrop against which to investigate the "gluing together" of the "local logics" of the system, in the sense discussed in Section 3.8d). We recall that "local logic" refers here to the "coherence" of various subsystems in relation to their respective proper environments. The questions are: (i) How should the local logics be glued together so as to maintain coherence for the system as a whole ? (ii) How does the system bring about and maintain the proper gluing?[7]

In the context of the nervous system of an organism (and depending on how complex an organism is being considered) the relevant structural "modules" (or, alternatively, structural substrates of functional "modules") could range in size from individual neurons to large populations. In this setting, and in the setting of autonomous systems more generally, one basic consideration (we we shall touch upon only tangentially in the present paper, in Section 6.4b)) will be to clarify the relationship of "modules" to the various components of the conceptual architecture such as "symbols", "representations" (cf. Section 5.7), and the "value system".[8]

---

6 In speaking of "approximate modularity" we do not mean to suggest that this is in some sense a flawed form of "strict" modularity. Rather, we are more inclined to regard the latter as a special "limiting case" of "approximate" modularity.

7 We wish to allow a good deal of latitude to the "module" notion. In particular, the modules in a particular "modular decomposition" need not all be at the same "organizational level". In addition, we may wish to consider multiple modular decompositions concurrently, while taking into account the various interfaces between them. (In this respect we are close to the modeling framework proposed in the *C. elegans* Working Paper cited in the footnotes to Chapter 1.)

8 An interesting approach to modular organization has been proposed by J. Mittenthal and his collaborators in the context of the organization of an organism (cf. the references given in the footnote to Section 3.2a). Their hypothesis ("the principle of matching") is that dissociable constraints on the organism will be met by dissociable clusters of processes, referred to as "dynamic modules". The modules mutually constrain one another via the necessity for "matching" at their "interfaces".

6.3 Finding relevant notions of "state"

The notion of "state", either as a primary or as a derived concept, figures prominently in models in the physical sciences, control theory, and computer science. We want to consider whether some variant or variants of the notion of "state" might have a basic role to play in the description or analysis of adaptive autonomous systems. We do not insist here on some sort of "canonical" state-space: On the one hand we do not adopt a reductionist outlook, so we are not concerned with trying to find a "base level" description specifying "all" the relevant variables. Moreover, we are sympathetic to the viewpoint of Willems (cf. Section 2.8) that state-space realizations may be highly model-dependent and non-intrinsic. However, even stopping short of a canonical state-space picture, it is still reasonable to ask whether there are notions of "state" which might be particularly suited to the setting of adaptive autonomous systems. In the present section we shall discuss some drawbacks to transposing standard notions of "state" to this setting. (This in effect recapitulates, from a somewhat different vantage point, certain of the difficulties discussed in Section 2B regarding the transposition of approaches from theoretical computer science and "abstract" control theory to the setting of adaptive autonomous systems). In the following section we shall examine one possible route towards relevant notions of state, namely via the value system.

We shall, as a starting point, assume that we have already been able to make some natural modular decomposition of the system, and shall consider two approaches to "state assignment", one emphasizing the modules themselves and the other the "communication channels" between them.

a) Assigning states to modules

We assume each module has its own "state space", with the state space of the system as a whole being the product of these "local" state spaces. For concreteness, we picture each module as functioning in a "Turing machine-like" fashion: for each input and each state, we give a formula for the output.[9] (Making provision for non-determinism would not affect the discussion to follow. We allow multiple input and output "ports" for each module, but this likewise does not really affect the discussion.). Among the drawbacks to this approach are the following: (i) Combinatorial explosion. Even in the extreme case of 2-state ("off" or "on") modules, there are 2^N states if N is the number of modules.[10] (ii) The "inputs" and "outputs" for the actual system can be extremely complex, context-sensitive, etc. None of this complexity is reflected in the above picture. (iii) In a related vein, the above approach is too "formal" in its treatment of the "inputs" and "outputs", regarding them purely as abstract tokens. Instead they should be represented as "symbols" (in the sense discussed in Chapter 7) having function and

---

9 While speaking here of "Turing machines", which are intended to carry out *terminating* computations, we want to give at least equal emphasis to "reactive systems" in the sense of Pnueli (see the second footnote to Section 2.13). Certainly, the systems we are interested in are submitted to a continual flow of "inputs" while producing a continual flow of "outputs".

10 Here we are concerned with "combinatorial explosion" as a problem for us in analyzing the system, arising as an artifact of *our* mode of "system decomposition" , rather than as a problem for the functioning of the autonomous system itself. However, if one assumes that the system functions algorithmically, any necessity for "search" might well pose problems for the system itself.

content. In particular, they can act by means of their specific physical embedding in ways not reflected in this "abstract token" picture. The above deficiencies in the representation of "inputs" and "outputs" are mirrored by corresponding deficiences in the representation of "states". (iv) The above approach does not throw any light the "logics" of the system (in the sense discussed in Section 6.2b)).

b) Assigning states to "channels"

In this approach one also specifies the state of each module and the rule of functioning of each module (which should be "simple"). However, in addition, one incorporates as part of the state of a module the state of each of its communication channels (i.e., the "messages" transmitted, or allowed/disallowed to be transmitted on these channels). This approach, in its stronger use of the analytic decomposition of the system into modules, is less susceptible to criticism(iv) above, but it does not circumvent problems (ii) and (iii). On the other hand, problem (i) is significantly mitigated as a result of the shift of emphasis away from the states of the modules to the states of the communication channels. (Even assuming "maximal connectivity", with every pair of modules communicating, this would replace the 2^N states noted above by N(N-1)/2).[11] However, an emphasis on the communication channels introduces new difficulties: (v) An adequate treatment of the "communication channels" would require greater attention to questions of time and timing. A couple of examples are: each module requires time to process its inputs; additional inputs can arrive while a module is still "processing" its current inputs. Moreover, for many purposes, thinking in terms of "channels" rather than "interfaces" between modules may be a gross oversimplification (cf. the final footnote to Section 6.2a)).

c) Summary

Our basic criticism of the above approaches to the notion of "state" is that these approaches are not really geared for embedded systems, let alone autonomous systems. They treat "states", "inputs", and "outputs" as abstract entities , and do not take account of their embedding-mediated "semantic content". Nor are they able to represent other than "formal" relations among these "entities". This means, in particular, that they cannot take account of "vague specifications" (cf. Chapter 4). Similarly, they do not take account of the (basic) constraints on the system, including the constraints associated with time.

6.4. "States" linked to the value system

a) "States" in a non-algorithmic setting

We have emphasized in earlier chapters the non-algorithmic character of adaptive autonomous systems. This is, in fact, the basic source of the difficulty in settling upon natural notions of "state" for this setting. Traditional notions of "state" are linked, either explicitly or implicitly to an underlying algorithmic picture, and it is not obvious that analytically useful notions of state can be defined otherwise. Put somewhat differently, "clean" definitions of "state" very likely are linked to correspondingly "clean" definitions

---

11 The approaches discussed in Section 2.11, in effect, circumvent combinatorial explosion of the state space by focusing on the "channels" (or histories of input/output tokens).

of "state transition". Additional "complicating" factors are the "emergent" character, and the dynamic vs. static nature of autonomous systems (in the sense discussed in Section 4.4c)). More generally, the difficulty in formulating appropriate notions of "state" may parallel the necessity for non-standard notions of "specification" (i.e., the notion of "vague specification") in the setting of autonomous systems (cf. Section 4.4). Just as "vague specifications" are not simply "blurrings" of sharp specifications, the relevant notions of "state" may not be simple "twists" of standard notions of "state". We wish to emphasize that "vague" is not equivalent to "stochastic" or "probabilistic", although there will be "vague specifications" where there is a concomitant stochasticity. This is akin to the point we make in Chapter 3 to the effect that "choice" should not be confounded with "randomness" (cf. the final footnote to Section 3.6b)(i)).[12]

---

12 We should perhaps be a bit careful in our formulation in the above paragraph. Certainly, an adaptive autonomous system can be viewed from multiple perspectives (and at various levels of detail), some of which are concerned only with "physical system" vs. "autonomous system" aspects. For such purposes physical-science models (or formal "computational" models) of traditional type may be what is called for. For example, "dynamical system" models of individual neurons, or of neuronal "circuits" incorporating various levels of biophysical/biochemical/electrophysiological/morphological/connectional detail surely are totally appropriate for many purposes, with the specific types of detail included depending on the purpose at hand. This includes highly schematic "neural network" models. Here, on the other hand, we are concerned with notions of "state" tied specifically to the system qua autonomous system. Our assumption, as put forward in earlier chapters (see. e.g., Section 3.6b)), is that the non-algorithmic modes of operation of the system will not be amenable to algorithmic analysis on our part.

As noted earlier, we are aware that this is an assumption rather than a tautology, a point which is illustrated by considering physical systems: Thus, motion under the influence of physical forces is "non-algorithmic" but may nonetheless be computable "algorithmically" based on equations describing the action of these forces. We are, in effect, questioning whether the analogues of such "equations" exist in the context of autonomous systems. Even if we are mistaken here, and algorithmic analysis is feasible for autonomous systems, it is still far from clear that the relevant "state spaces" would correspond in any direct way to those linked to particular "physical system" models associated to the autonomous system. In a sense, this question of "direct correspondence" between "physical system states" and "autonomous system states" comes back to the well-worn distinction between "reducible to" and "supervenient upon". There is, however, an additional factor in the present setting which distinguishes it from the usual contexts in which the above distinction is discussed: In the present setting, to say that "descriptive level" A (namely that of "autonomous system states") is supervenient upon "descriptive level" B (namely that of "physical system states") tacitly assumes that we, in fact, are in possession of a descriptive level A. However, the issue at hand is, precisely, to construct such a descriptive level, presumably using a vocabulary linked to the structural/functional architecture of the autonomous system. Actually, in the present setting, to suggest that there is a single, "uniform" descriptive "level", drawing upon a single, uniform set of descriptive "primitives", may be misleading. Our suspicion is that, due to the very nature of autonomous systems, any description of the system (qua autonomous system) initiated at one "level" will inevitably "propagate" to other levels.

It is perhaps the case that the relevant "state spaces" would have the property that "vague specifications" can naturally be "formulated" in them , whereas such specifications can perhaps not be naturally formulated in the "physical system" models. Perhaps a corresponding point can be made regarding the notions of "behavioral equivalence" or "equivalence relations"; cf. Sections 2.10 and 7.3c)(iv). An interesting case in point here may be the differences between performing a spectral analysis of a speech waveform vs. an analysis into linguistically relevant features, or "phonemes". (In using the term "formulate" here, we do not want to suggest a "propositional" formulation. In fact, this may be an instance where an "internalist" is essential: rather than speak in terms of how a vague specification is "formulated" or "expressed" in abstracto, perhaps one should refer only to the act or process of "specifying", as carried out be the system itself; cf. the distinction we make in Section 7D between "categories" and "categorization". As an amplification of this point, perhaps it is the case that the natural state spaces can only be approached from an internalist perspective, whereas the "physical system" models are inherently linked to externalist perspectives.)

Despite the potential pitfalls noted above, it may nonetheless make sense to keep these concerns in the background for the time being, and to look for relevant notions of "state". The question may be rephrased as: what kind of information should a mathematical or a logical definition of a "state" incorporate? Giving an explicit answer may be quite difficult, but we are prepared to make a guess as to an implicit partial answer. It is natural to conjecture that one set of relevant state variables should be the variables in terms of which the activity of the value system (both in its "evaluative" and "biasing" roles) can be expressed. We have in mind here the higher-order value system as well as the primary value system associated with the fundamental goal of the system: to maintain itself as a system. In connection with the primary value system certain of the relevant state variables would be those linked to the basic constraints, a case in point being "the" metabolic variables.[13]

b) "Localization" relative to a module

In keeping with our anti-reductionist outlook, we do not think that there is a natural set of "basic" state variables in terms of which all the others can be expressed. On the other hand, we do not think that the remedy for this is to attempt to assemble "all" the relevant state variables into a single macroscopic "state vector". The relevant variables form too heterogeneous a collection, associated with too many distinct organizational "levels", for such a list to be useful.[14] Moreover, such a list would not of itself facilitate the analysis of the manifold relations among the state variables. What is called for, we think, is to look for approximate "modular decompositions" of the system that in some sense correspond to (or facilitate the discovery of) approximate "modular decompositions" of the value system.[15] From this vantage point a "good" modular

---

In the "concurrent constraint programming" approach of Saraswat (cf. the footnote to the introductory paragraph of Section 2B), the picture of the "store" (or "memory") as valuation is replaced by that of the store as family of constraints (where, roughly speaking, "constraints" in Saraswat's sense includes both "constraints" and "specifications" in our sense. We find this an attractive shift of emphasis. Despite the fact that this approach is formal vs. embedded (so that the "constraints" are formally expressed vs. physically realized), and despite the considerations raised above, we suspect that this idea may, in some form, be transposable to our setting. (It is perhaps interesting in this regard that one of the motivation's of Saraswat's approach is to be able to deal with "partial information", or "partial specifications". One of the tasks associated with a transposition to our setting would be to determine how to deal, instead, with "vague specifications"; cf. Section 2.10.)

13 In looking for relevant state variables we do not wish to be dogmatic in restricting attention exclusively to the value system. For example, we surely want to take account, at some level, of the processes of synaptic modification linked to learning and long-term memory, and which are at the root of the self-modifying circuitry of even the most primitive nervous systems. It is not clear to us that the relevant variables here are necessarily linked directly to the value system.

14 These "levels" need not be arranged, or "layered", in a straightforward "hierarchy" (although there may well be "local" layerings of this type). Nor can one assume that one could restrict attention to state variables of uniform "kind". An example of a "kind" distinction (borrowed from "computer functionalism") is that between the "physical" states of a computer and the "functional" states associated to the computation being performed.

15 This is slightly reminiscent of the "matching principle" of J. Mittenthal (see the second footnote to Section 6.2b)). Of course, one may make the reciprocal argument: If one has a "modular decomposition" which is natural on whatever grounds, then this may point towards relevant state variables, linked to the value system or not.

decomposition of the system would be one which is consistent with the "partitioning" of a particular subset of the state variables into subsets of variables "localized" to (or naturally "belonging to") individual modules, and interacting only via the module interfaces. More succinctly, we want modular decompositions which localize various of the state variables, whether these state variables are linked to the value system or to other facets of the autonomous system's architecture (e.g., to such other components of the "conceptual architecture" as are discussed in Chapter 7).[16]

B. Control:Maintaining Coherence

We have argued in preceding chapters (cf. Sections 2.6d), 3.4, 3.8d) and 3.8e)) that in the context of autonomous systems the primary question of "control" is how the system achieves and maintains "coherence" of its overall behavior, rather than how it succeeds in carrying out specific tasks or achieving specific goals. This point of view is consistent with our "internalist" emphasis, together with our emphasis on the "emergent functionality" of the system and, correspondingly the emergent character of its control organization. In the present section we wish to continue our discussion of control from this perspective. We shall, in particular, use the picture of "interacting modules" as a convenient backdrop against which to examine several modes of control, including control making essential use of "symbolization" (in the sense of Section 5.7).

The "modular decompositions" discussed here will be based on functional rather than on structural criteria. Thus, we shall speak in terms of "sensory" modules, "motor" (or "action") module, and "control" modules.[17] To forestall misunderstanding we wish to remind the reader of the points made in Section 6.2 regarding the distinction between functional and structural decompositions as well as the dynamic vs. static nature of modular decompositions. In particular, in speaking of "control modules" we do not mean to suggest that these are necessarily realized as separate physical structures or "devices" localized externally to the systems being controlled. In addition we wish to add another couple of comments here. First, the preceding terminology should not be taken to suggest

---

Cf. , also, Rowley, S.G. and C. Rockland (1991) The design of simulation languages for systems with multiple modularities, *Simulation* ,.56:3, 153-163.

16 We speak of partitioning a "particular subset" of the state variables since we think it highly unrealistic to expect that a single modular decomposition (other than the trivial one in which the whole system is regarded as a module) will allow the localization of all the state variables. It would not, in fact, be surprising if certain of the state variables are not in any useful sense "localizable". By "partitioning" we also include, for example, the "localization" to spatial subdomains of a single variable defined over a global domain.

One class of state variables which are likely to be highly localizable (for a wide range of modular decompositions) are the metabolic variables. Indeed, in many cases it is not altogether fanciful to postulate that the "requirement" on the autonomous system "to maintain a state of relative adaptation with the environment" (cf. Section 3.1) passes over to individual modules, with "proper environment" substituted for "global environment". We are thinking here of cases where the individual modules are in some sense "quasi-autonomous". In such cases there may well be "local" metabolic variables for the individual modules.

17 It may be useful to bring a broader construction to the term "action module" than identifying it exclusively with "motor module", so as to accommodate the actions of a subsystem on its proper environment. Thus, "control" modules (and perhaps even "sensory" modules) may in many instances be viewed as "action" modules from this standpoint.

a simplistic input/output "flow diagram" of the form: sensory modules --> control modules --> motor modules . By the same token, neither should "sensing" be interpreted as a purely passive, receptive process. For example, the sensory apparatus can function in a "goal-influenced" or "attention-influenced" exploratory fashion, with its particular "sensitivities" (including its "receptive fields") and modes of operation being dependent on history and context.[18]

6.5 The need for control

a) Resource limitations[19]

As discussed in Sections 3.2 and 4.2, the system, in the process of meeting its "fundamental goal", is subject to various limitations of both external and internal "resources", among them limitations associated with the "basic constraints" on the system. For example, the system must interact with a rich, complex environment while having available only a limited number of sensory and motor modules. This implies, in particular, that there will be a need to coordinate the "spatio-temporal" access of the various "goals" or "specifications" of the system to the requisite resources for realizing them, notably the "action" modules, which are the main resources of the system. Even if no special coordination were incorporated, the system, being embedded, would "do something". However, this behavior could well be "incoherent" and even self-destructive. It is to avoid such incoherence that "control" modules need to be introduced.[20]

---

18 Somewhat more generally, we regard the tripartite division into "sensory", "motor", and "control" modules as a convenient basis for discussion rather than an intrinsic or a priori functional taxonomy (or as a set of functional "building blocks"), however "coarse-grained". (Thus, for example, while it makes good sense to ask how different "forms" of memory are realized, it is surely not a substantive question to ask precisely what kinds modules of the above type are "involved in" memory. The answer would probably have to be: "All three". To speak of the resulting concatenations of modules as "memory modules" would, of itself, be nothing more than a terminological convenience.) In fact, we think that, in a more formal treatment, individual functionalities will not be "definable" in isolation, but will need to be defined in relation to one another (as is the case with a set of "axioms" defining the primitives of a formal theory), as well as likely requiring reference to other "levels" (e.g., structural) of the overall system architecture; cf. the reference to "propagation" from one descriptive level to another (in the first footnote to Section 6.4a)). Similarly, one may expect that function (in the sense of "role") must ultimately be referred to the value system (which, reciprocally, must be referred to the structural/functional architecture of the system). The foregoing remarks do not vitiate attempts at modular decompositions, but merely point out that the definition of such a decomposition must incorporate a specification of the interface relations among the modules.

19 In the context of formal semantics for real-time concurrent computation, Joseph and Goswami have emphasized the necessity for the semantics to take into account the constraints on program execution imposed by resource limitations. (See, for example, M. Joseph and A. Goswami, "Semantics for specifying real-time systems", pp. 272-275 in C. Rattray (ed.), "Specification and verification of concurrent systems", Springer, 1990.)

20 The preceding discussion is deliberately oversimplified in order to bring out the main point. However, to forestall misunderstanding, it may be advisable to make a few clarifications: (i) In examining how coherence is achieved it is necessary to distinguish carefully between the "evolutionary", "developmental", and "somatic" time axes. Our discussion in this chapter is primarily directed at the issue of how coherence is maintained by the system in "somatic" time, once evolution and development have already "done their jobs". This to some extent accounts for our speaking as if the sensory and motor modules exist "prior" to the control modules, although even in the setting of "somatic" time this is surely a significant oversimplification. (One "image schema" motivating part of our discussion is that of "synaptic

b) Control as the "gluing together" of local logics

A related perspective towards control is the "gluing" picture, discussed in Section 3.8. Here the overall system is conceived of as being "glued together" from various smaller "parts" or "subsystems" which, regarded individually, have their own coherent "logics".[21] This "gluing" may take the form of various logical or material overlaps or interfaces associated with the various constraints or specifications on the system as a whole. For example, two such subsystems may be "glued together" via the sharing of common motor or action modules. As noted above, the "local" logics of the individual subsystems are coherent. But these logics also need to be, or must be made to be, compatible. The question of "control" may then be viewed as how to "glue together" the local coherent logics so as to obtain global coherence for the "logic of the system as a whole. It may require the introduction of additional "control" modules to carry out this "gluing", or to serve as the "glue".[22]

This should be regarded as a dynamic vs. a static process. That is, not only is it necessary to obtain, but to maintain coherence. For example, the particular "gluing" necessary for global coherence may be time- and context-dependent. Nor should the subsystems themselves be viewed as static, although the subsystems may deliberately have been selected (by the analyst) so as to be "stable" on a longer time-scale than the "gluing" between them. It may, in fact, be necessary to "work" to maintain the local coherence of the individual subsystems.[23] Moreover, it must be emphasized that all this work must be done by the system itself. That is, part of what constitutes "coherent" activity by the system (as, presumably, "evaluated" by the value system) is the

---

strengths" becoming modified over time, in conjunction with the "experiences" of the organism. A basic question is how such modification is done "coherently"). (ii) In speaking as if the "goals" and "specifications" of the system must "compete" for resources (including memory "storage" and "access")we do not mean to suggest that they necessarily exist prior to and independent of the available resources. As we have stressed in Section 3.3 the system's explicit outward-directed goals (in contrast to it's implicit, self-referred "fundamental goal") are conditional concomitants of many factors of history and context, including the available "resources". Putting this very crudely, the system must "learn" (on whatever time axis) to want what it can get. This is one aspect of "coherence". (iii) The necessity for coordination should perhaps be viewed from a wider perspective than consideration of resource limitations. For example, certain behaviors may, via their outcomes, destructively interfere with or vitiate one another even though they do not make use of the same motor "apparatus". (To take the example of a simple organism, locomotion may interfere with feeding.)

21 It may, in some instances, be convenient to regard these "parts" or "subsystems" as "modules" in their own right, though not infrequently we would regard them as themselves having modular decompositions, and as interfacing to one another via common, or "shared" modules.

22 We emphasize once again that these putative "control" modules may, but need not, require the introduction of new "structures" as contrasted, say, with modification of "connectivity". (Indeed, it is too soon to say precisely how useful this "module" terminology will prove to be in the present setting of control.)

23 For example, certain of the individual subsystems may themselves consist of yet "smaller" subsystems, themselves "glued together" via "control" modules. However, we distinctly do not wish to emphasize a "hierarchical" picture of "gluing".

Just as we are assuming the constituent subsystems to be more "stable" than the particular "gluing" that links them, we also assume that the burden of global coherence can be borne largely by this gluing, without the additional need for substantial internal modification of the individual subsystems. A priori this is not obvious. Indeed it is even conceivable that global coherence may require a degree of local *incoherence*. After all, once "glued together" the individual subsystems no longer function in isolation.

maintaining of "coherent gluing" (again, as evaluated by the value system). In particular, the structural/functional architecture of the system must provide the substrate for this.

6.6. "Principles" for control design

In this section we shall examine control from the point of view of a system designer. Assuming one were to seriously entertain the prospect of designing an (artificial) adaptive autonomous system, are there principles or guidelines that one might naturally seek to impose on the control design? If so, one might also seek to ascertain whether these principles are, in fact, realized in the "control architecture" of biological organisms. We shall raise three such "principles" for consideration here. These will form a basis for the discussion in the sections to follow, in which we shall consider several modes of control, both in the context of our putative artificial system and in the context of biological organisms. We readily acknowledge that the putting forward of abstract principles is no more than a preliminary step to showing how they can actually be realized.

a) <u>"Gluing" local logics in design time</u>

During the initial, "conceptual" phase of design the designer will typically begin by trying, on a partially intuitive basis, to "break down" the initial specifications for the system into specifications for "simpler" component subsystems or modules, which are to be assembled to form the system as a whole.[24] This decomposition will be influenced by what types of "parts" happen to be readily available. Thus, the resulting set of subsystems will reflect a mix of "off the shelf" and "custom" design. In any case, these subsystems have their own rules of functioning, which can be rather complex. During this phase of the design more attention may be given to the specifications for the subsystems than to those for the interfaces between them. As a result, when these subsystems are put together the system as a whole may not fulfill the intended specifications, even though the various individual subsystems may fulfill the specifications set for them (at least prior to being put together). Indeed, many conflicts and incoherence may arise, so that not only are specifications not met, but wholly "unintelligent" behavior may result. Correction of the "problem" (which, typically, will be an iterative rather than a one-step process) may, in

---

24 Our discussion here is not intended as anything more than a highly simplified and caricatural depiction of the design process. To some extent what we are doing is transposing our previous discussion regarding "gluing" from the context of "somatic" time to the context of "design" time.

For purposes of our discussion the term "subsystem" may be interpreted in various ways. Two extremes are: (i) A "horizontal" decomposition into particular "functionalities" which are "incomplete" in themselves, and must be integrated to yield "behavior". (ii) A "vertical" decomposition into individual "behaviors", each "coherent" in itself, which must be properly coordinated to produce the desired over-all system behavior. (This is one variant of Brook's "subsumption architecture" picture; cf. R. A. Brooks, "A robust layered control system for a mobile robot", IEEE Journal of Robotics and Automation, Vol. 2, Number 1, pp. 14-23, 1986.) Part of the initial motivation underlying this mode of decomposition is that it allows the designer to replace, or "subsume", an existing "simpler" behavior by building a more complex behavior "on top of it" via appropriate interfacing with the "modules" producing the simpler behavior. This latter aspect of "subsumption" bears a curious resemblance to ideas of John Hughlings Jackson (1835-1911), one of the founders of clinical neurology, regarding the "evolution" and "dissolution" of nervous system function; cf., J. Taylor, ed., "Selected writings of John Hughlings Jackson", vols. 1 and 2, Hodder and Stoughton, London (1931/32); reprinted, Basic Books, New York, 1958).

the worst case require total reconsideration of the initial "break down" into subsystems, but in more fortunate circumstances limited modifications or enhancements of the subsystem specifications may suffice. In the best of cases the subsystems can be left more or less unaltered, with the burden placed instead on the redesign of the interfaces among these subsystems. perhaps together with the introduction of additional subsystems or modules. As in our discussion above of the "somatic" time context (cf. Section 6.5b)), it may be convenient to regard this as the introduction of "control" modules to "glue together" the logics of the various subsystems.

In the context of evolutionary biology there is no "designer", nor are there preset specifications in the above sense (cf. Chapter 4), to be "broken down" into subsystem specifications. However, there are perhaps some relevant parallels to be drawn. Evolution does not start from scratch, but begins from a base of "apparatus", or "stuctures", or "functional subsystems" already at hand, and "glues" them together in new ways, perhaps modifying them or adapting them to new "roles" in the process. This "gluing" may also involve the introduction of additional "modules". Certainly, those systems in which the "local logics" are glued together "incoherently" will be at a selective disadvantage. Indeed, in some cases adaptive "success" will require that the system be endowed with the capacity for ongoing modification, during somatic time, of the "gluing" of its local logics so as to maintain coherence. An implicit concomitant of this is an internal value system to serve as an "arbiter" of coherence (or of deviation from or approach towards coherence).

b) Principles for control

The above picture of "coherent gluing" provides a backdrop for discussing some putative principles, or guidelines, for control design for adaptive autonomous systems:

(i) Emergence

There should be a "division of labor" between the designer and the system itself. The designer should not attempt to make "decisions" in design time that require information available only in somatic time.[25] Rather, the designer should "build into" the system the capacity to modify its own control architecture and the capacity to learn how to construct and choose between different control "strategies". It is in this way that the system will be able to adapt to the contingencies of somatic time (see Section 6.10). In particular, the "full" control architecture should be emergent over somatic time.

(ii) Conservatism

To the extent possible the designer should, like evolution, not design from scratch, but should instead make use of, and build upon, existing subsystems (say, drawn from earlier designs) which individually already function coherently. In keeping with this "conservative" viewpoint, the designer should not seek to modify the *internal* architecture of these subsystems, which may be quite complex. Instead, the emphasis of the design, apart from selecting the relevant subsystems, should be on devising the proper external interfaces. This is not to say that some internal modification will not be needed.

---

25 We do not regard "adaptive control", as the term is typically used in the field of control theory, as transferring enough responsibility (or, alternatively stated, as providing enough latitude) to the system itself. Here the designer provides a preset parametrized family of control laws. Then the system, at run time, selects the proper parameter values, based upon criteria also preset by the designer.

However, in keeping with principle (i) above, this task should, to the extent possible, be carried out by the system itself, in somatic time, without the intervention of the designer. That is, the individual subsystem architectures should be "compliant" rather than "rigid", and should be able to modify themselves as necessary in somatic time, possibly making use of learning processes. [26]

We have discussed "conservatism" in the context of design time, but we think that the same principle, including the building upon existing subsystems, can be transposed to the setting of system self-modification in somatic time.

(iii) Opaqueness of modules[27]

Neither the design of the "control system", nor its functioning during somatic time should depend on the designer (resp., the control system) having detailed "knowledge" of the internal structural/functional organization of the "modules" which are controlled. The only "information" that should matter is that obtainable via interaction with these modules through their interfaces.[28]

---

26 In the picture we have in mind these modifications will be comparatively circumscribed. However, this may be too simplistic a view, especially if one also wishes (as we do) to consider biological organisms, where the "developmental time axis" must be taken into account. In fact the associated genotype-phenotype distinction brings to the fore the following basic consideration: In speaking of incorporating existing subsystems largely intact one must be careful to distinguish between "slight perturbations" of the subsystem architecture and perturbations of the "specifications" (sharp or vague) for the subsystem's "assembly". Even if the assembly specifications and the subsystem architecture each possess a degree of "modularity", the two distinct "modular decompositions" need not exhibit any straightforward homologies. (Indeed, the subsystem itself may not correspond to a "module" in the assembly of the overall system.) This is particularly to be expected if the subsystem is the result of physical interactions of various elements under the "orchestration" of developmental "programs" or biases.) Thus, a small perturbation in the (physically realized) assembly specifications could result in major alterations in the structural/functional architecture of the subsystem. A case in point is provided by the notion of "heterochrony", whereby small perturbations in certain "control genes" can result in modification of the timing "structural gene" activity , so as to yield major modifications in the architecture of the system constructed; cf. S. J. Gould, "Ontogeny and phylogeny", Belknap Press (Harvard U.), 1977.

Our presentation both of the principle of "conservatism" and of the principle of "opaqueness of modules", to be discussed next in the text, may suggest that what we have in mind are simply variants of the notions of "design for reusability" and "modular design" which are already being emphasized in engineering design. However, there are at least two differences: (i) The engineering design approaches emphasize "formal specification" methods, with *sharp* specification of module interfaces and of externally observable module behavior (i.e., observations of interactions at the interfaces). We, by contrast, wish to emphasize *vague* specifications. (ii) Similarly, the corresponding engineering design approaches do not emphasize "compliance" and "self-modification" of the subsystems.

27 What we are calling "opaqueness" is sometimes referred to as "transparency" in the context of computer science.

28 It is easy to see why this principle is relevant to the context of engineering design, not least in connection with software reuse and modification. It encourages the use of appropriate high-level abstractions and, in the process, facilitates the communication and coordination of the efforts of different design teams concerned with different subsystems. In addition, it impedes "destructive interference", whereby one design team introduces "hidden" design modifications which effect the work of others. In the context of organisms this "principle"(to the extent that it is "used" by the system itself rather than just by the analyst seeking to describe the system) must be interpreted more broadly, inasmuch as neither "modules" nor their "interfaces" come labeled as such.

6.7 Current approaches to design of control laws: state-based and behavior-based models

In this section we shall discuss two distinct methodologies currently used for the design of control systems. In the first instance, where accurate models of systems are available from physical laws, such as in the control of rigid bodies or the attitude control of a satellite, the evolution of the system is described by ordinary differential equations relating internal "state" variables x (such as position and velocity) to "control" variables u (such as force). This set of equations may be regarded as a dynamical system in x parametrized by the control function u (viewed as a function of time). Certain functions of the state variables are designated as "output" variables, and are denoted by y. Typically, these are variables which can actually be measured. Thus, the choices of u and y correspond, respectively, to a choice of sensors to carry out measurements and a choice of actuators to carry out control actions. The sensors and actuators may themselves have dynamics, which then needs to be modeled. It is assumed that this dynamics has been incorporated into the state-variable description of the system. The objective of control is to choose u so that the system exhibits some specified behavior. For example, one may wish the state trajectory to follow some desired path, subject to constraints on the control (e.g., energy constraints) and possibly on the state. In the case of feedback control, u should be determined as a function of the past of the measurement y. Very often, as in problems of guidance and control of aerospace vehicles, the determination of u can be formulated as an optimal control problem, and can be solved using calculus of variations or dynamic programming. Process noise and measurement noise can be accommodated by formulating the problem of control in the context of stochastic dynamical systems and then solving using stochastic dynamic programming.

Now, almost by their very definition, models of this type satisfy principles (ii) and (iii) of Section 6.6b), with the control interface provided by the parameters u. They do not satisfy principle (i), but this hardly seems a drawback here, since the models are already so simple (conceptually, even if not mathematically). However, this is precisely the point. Such representations of a system, via explicitly given dynamics linking state-variables and control-variables, and with sharply specified control interfaces, will only be feasible in comparatively "simple" contexts, typically where the predominant factors are physical laws. Situations where accurate underlying models (based, for example, on physical laws) are difficult to obtain are generally not suited for treatment via such internal state-based approaches. A key difficulty is that model uncertainties cannot readily be captured in parametric form in a state-based setting. For example, a basic desideratum is that small changes in external behavior should lead to only small changes in the state description, in particular in the dimension of the state vector. However, in most situations the dimension of the state vector may change drastically, even though the input-output behavior changes by only a small amount. One approach to this difficulty is via the methodology of network theory[29] : Approximations and interconnections are done externally, and internal

---

29 Cf. Section 2.6e) and the reference in Section 2.8 to the behavior-based approach to control developed by Willems. We note that, as used in the present section, the term "internal state" does not have any connection to a "value system".

state-based descriptions are constructed only after these approximations and interconnections have been performed.

An example of a situation where model uncertainty is considerable is the control of a macroeconomic system by a central bank. The process to be controlled is the dynamics of the economy. The "control parameters" that the central bank can influence directly are various rates of interest and government bond offerings. It is far from clear that internal state-based models of the above type can, even in principle, be used to represent the system here.The same is very likely true for the control processes of a large industry or of a large company.[30]

6.8 Control of communicating modules

We shall next touch upon two modes of control in the setting of communicating modules:

a) Control via transition restrictions [31]

Our setting will be a highly stylized variant of the scheme discussed in Section 6.5. We assume we have a system consisting of communicating "modules" which, for simplicity, we shall view here as finite automata (with multiple input and output channels). To the state-transitions of these automata there are associated "output actions". These may consist either of actions on or in the "external environment" (e.g., by "motor" modules; this includes locomotion), or of "messages" sent as inputs to other modules. We also assume that the system is supposed to satisfy certain specifications, perhaps linked to some "fundamental specifications" on the internal state of certain of the modules (e.g, "metabolic" modules) or, alternatively, specifications on a subset of the state variables (e. g., "metabolic variables") of many of the modules. There are constraints (e.g., the "metabolic" variables tend to "run down") which imply that the system needs to act (e.g., obtain "food") in order to satisfy these specifications. (In addition, there are constraints linking system actions to changes in state variables; e.g., linking food intake to alterations in metabolic variables). As in Section 6.5, we are concerned with coordinating the "output actions" or state-transitions of the various system modules so that the system specifications are met. This includes, in particular, coordinating access to common "resources", including other modules.

---

30 In the present context it is not quite clear how to discuss the principles of Section 6.6 b), inasmuch as there are no obvious choices of "subsystems" or "modules" in which to frame such a discussion. However, we suspect that principle (i) is probably relevant. Given the "control parameters" available, principle (ii) is perhaps satisfied by default. It would certainly be a mistake to take too simplistic a view towards principle (iii), and attempt to interpret economic variables as simple macroeconomic aggregates, to be regarded as "inputs" or "outputs" of modules.

31 We have actually worked this approach out in some detail in an earlier set of (unpublished) notes. We shall not, however, do more than give a sketch here, since our conclusion is that the approach is too weak to be of use in connection with adaptive autonomous systems. In particular, we shall say nothing about timing, buffering, additional module "parameters", etc.

The Wonham-Ramadge approach to supervisory control of discrete-event processes (cf. Section 2.7) also is based on the use "transition restrictions". However, both their approach and their problem setting are quite different from those discussed in the present section. In particular, in their setting, what constitutes desired or acceptable "plant" behavior is sharply specified in advance.

One can attempt to build a control system by introducing additional modules, i.e., "control" modules, whose "output messages" to other modules consist of lists of allowed transitions. Upon receipt of such a message, the recipient module can only choose a transition from among those on the list. This introduction of additional "control" modules may be iterated several times, resulting in successive "layers" of control modules. The organization is not "hierarchical" (or "tree-like"), however. Messages, including "allowed transition" messages, can also be sent from "lower-order" to "higher-order" modules, as well as between modules in the same "layer". The overall system may be organized into (possibly overlapping) "subsystems" of modules.

This approach to control, at least as it stands, has several limitations which render it inadequate for adaptive autonomous systems:(i) Perhaps the most serious is that it can give rise to logical failure. A given module might be subject to control by two control modules which send two disjoint lists of allowed transitions. In this case the module would be unable to make any transition (even the "null" transition which consists of doing nothing). So, in attempting to resolve one set of potential conflicts (over access to "resources", one may be introducing new conflicts. It is not clear that this problem of logical failure can be resolved by adding successively more "layers" of control. In particular, the process may not yield "closure" after a finite number of iterations. (ii) Moreover this mode of control is at variance with control "principles" (i) and (ii) of Section 6.6. All the real work must be done by the designer, without any somatic time "emergence" of control organization. Nor does the approach appear to offer any particular guidance to the designer. Also control modules would need to have a detailed knowledge of what is going an inside the modules or subsystems they are supposed to control. This may be expected to lead to combinatorial explosion problems as soon as the modules or subsystems reach any degree of complexity.[32] (iii) This approach does not provide the various modules with "choice procedures" for selecting from among allowed transitions. One could leave this choice up to chance, but this seems highly unnatural.

b) Control via subsumption

A different mode of control, focusing more directly on control of communication channels, is related to a variant of Brooks' "subsumption architecture" for autonomous robots (cf. the footnote to Section 6.6a)). More specifically, this mode of control is associated with a technique of "incremental design", by which the designer may replace a given "hard-wired" behavior (or "level" of behaviors) with a more "sophisticated" level of behavior (intended to add new functionality or to compensate for deficiencies in the original level), without having to carry out a major ab initio rewiring of the overall system. The idea is to layer the new behavior level on top of the old one, in effect drawing on the resources used to generate the original behaviors. For example, one may begin with a level n°1 robot design (say, a "wandering" device) which works perfectly

---

32 In the present setting, where the control is not "emergent", these would be problems for the designer. To get an idea of the kinds of problems that can arise here, it may be worthwhile to consider analogous difficulties in the Wonham-Ramadge theory . In the case of incomplete observations of the plant to be controlled, the problem of supervisor design is intractable in the sense of computational complexity theory. (See J. N. Tsitsiklis, "On the control of discrete-event dynamical systems", Math. Control Signals Systems, 1989, 2: 95-107.)

well in a certain environmental contexts (for example, where there are sufficient uncluttered spaces, and where the robot is idle and has sufficient power reserves). But perhaps the robot does not function properly, in the sense that it tends to bump into walls, or in the sense that it does nothing useful. Then one may attempt to incorporate additional behavioral competences (e.g., such as inhibiting wandering and putting the "motor modules" to other use when obstacles are detected) in a succession of "higher-level" designs layered on top of one another.

In this setting a "behavior level" may be realized via a family of communicating modules (among which we may include the "environment"). The method of "layering" or "control via subsumption" (as we wish to use the term) may be described as follows: We start with two communicating modules , say $M_1 \rightarrow M_2$. In control via subsumption another module, say N, associated with a different behavioral level , may either suppress the output of M_1 towards M_2 , or may replace the message sent by M_1 to M_2 with its own message. This may be represented by the diagram: $M_1 \rightarrow M_2$. Here it is not M_1 whose

$N \rightarrow \uparrow$

function is modified, but the communication channel between M_1 and M_2.

This control procedure satisfies principles (ii) and (iii) of Section 6.6, this in some sense being the main point of the procedure. (Of course, in actual practice, this procedure, as well as these two "principles", must be interpreted as "guidelines" for design rather than as rigid rules.) Just as clearly, principle (i) of Section 6.6 is not satisfied. This is part and parcel of the difficulties with "behavior-based" approaches, incorporating predetermined priorities among behaviors. We discuss our objections to such approaches in some detail elsewhere in this paper; see, e.g., Sections 2.10, 2.14, 5.4b) and 7.11a).

We wish to conclude this section with two observations regarding the behavior "levels" which figure in the above picture. First, lower "levels" cannot control higher "levels". Second, the various distinct "levels" are really of uniform type, all of them being generators of behavior in the external environment. They do not correspond to different levels of abstraction or representation.

6.9 Control of interface relations via symbolization

From this point onward our emphasis shifts back to biological vs. artificial systems. In keeping with this emphasis, we shall freely make allusions to "neural systems". However, unless we explicitly state otherwise, our use of such "neural" terminology is intended in a metaphorical rather than literal sense. We are not attempting here to make assertions about the actual neural organization of specific organisms. (In particular, we don't wish to make claims as to the existence or character of specific "neural modules").

Our discussion in the present section will deal with a special case of the "control as gluing" picture: namely, the situation where two modules M, N each have interfaces to the same module E. The problem we shall consider is the coordination of "access" to E (regarded as a "shared resource") via control of the interfaces to E. We shall contrast two modes of control at opposite extremes of "directness". The first ("reciprocal inhibition") is a particular mode of "direct control", where no additional "control modules" need to be

introduced. The second is an indirect mode of control making essential use of symbolization.

a) Direct control of interface relations

We start again with the picture of "control of interface relations " : $M_1 \rightarrow M_2$, $N \dashv \uparrow$ discussed above in connection with the "subsumption architecture". Here N is a module which controls the communication or interface relation between M_1 and M_2. However, we want to work at a more "abstract" level than in the preceding section. In particular, we do not want to insist that N be a component of a " behavior generator", nor that N be at a "higher level" than M_1 or M_2.

Perhaps the most basic means of "conflict resolution" within neural systems (both internally and at the periphery, where the systems interface with the environment) is "reciprocal inhibition". We shall transpose this notion to our more abstract setting. It is represented schematically in Figure 6.1, via the symmetric arrows drawn between the two modules M and N both having access to the same module E. Depending on the relative strengths (and on the timing) of the reciprocal "signals" passing between M and N, one of them inhibits the other and "gains control" of E.[33] Notice that here coordination of access to E or, alternately phrased, the control of the interface relations between M and E (resp., between N and E) is achieved without the intervention of a higher-level "controller". We shall assume that control via reciprocal inhibition can arise in two ways: as part of the "pre-wiring" of the system and also via "circuitry modifications" linked to learning (in somatic time).

b) Control via symbolization

As we have seen in Section 5B, it may be advantageous for the system to be able to build a symbolic environment in which to experiment before committing to specific courses of action in the actual environment. We illustrate such a use of symbolization in the above setting, in which there are two modules M, N both having access to the same module E. We shall assume, for concreteness, that E is an "effector" module, which carries out action, motion, etc.... As an alternative to having the control of E be determined by the direct competition of M and N (via reciprocal inhibition), it might be preferable if the "decision" could be based on the anticipated results of the outcome. That is, it might be preferable if the system were able to "project" plausible courses of action by E in the two alternative cases where M or N had control, and were able to test the effects of these courses of action on an "internal representation" of the world. This process could be realized by, or regarded as, a "higher-level controller" which constructs a representation of the environment and acts symbolically on this representation. Based on an evaluation of the results of the symbolic actions, the controller decides to "give permission" to one of the actual actions and to inhibit the other. Thus far we have regarded E as an effector module acting on/in the external environment. But E could

---

33 Lest our reference to "signal strengths" and "signal timing" obscure this point, we wish to stress that our "modules" and their "communication channels" or "interfaces" are intended to be taken at a fairly abstract level, and not (for the present) associated to specific structural/functional substrates. For example, we are not attempting to identify "channels" with axons, or "signals" with sequences of action potentials.

equally well be a different type of module, not at the periphery of the system (i.e., not directly interfacing with the external environment). While not acting on the external environment, E nonetheless acts on its own proper environment (cf. Section 6.2a)), perhaps via quite abstract neural actions. The "strategy" of symbolic action on an internalized symbolic world carries over equally well to the present setting. The only difference is the higher level of abstraction: the world being symbolically represented is the world of a module (or subsystem) inside the system rather than the external world of the system. Such symbolic action (together with its evaluation) forms a basic component of "planning".

c) Control via symbolization: reprise

We wish to give a slightly more detailed (though still rather crude) formulation of "control via symbolization". As above, we begin with two modules M and N, having access to the same module E, which acts on its proper environment. The "controller" consists of a set of processes which:

(i) Have available (and if necessary, construct) a symbolic representation of the proper environment of E.

(ii) Carry out symbolic actions on this symbolic representation "mimicking" the actual actions of E on its proper environment, corresponding to the two alternative cases when M or N has control of E. (The determination or "estimation" of what the actual actions of E would be in each of these two cases might itself involve symbolization processes.)

(iii) Evaluate the results of these symbolic actions on the symbolic environment using symbolic values.

(iv) Compare the results of these evaluations. (Actually, there may not be a clean separation between the processes of evaluation and comparison. In any case, it will likely be necessary to provisionally "store" either the results of the symbolic actions, or the evaluations of these results, in some form of "working memory").[34]

(v) Give permission to M or N and inhibit the other.

The primary problem with the preceding picture of "control via symbolization" is not its high degree of abstraction; indeed, we shall have to deal with much worse in Chapter 7 ! Rather, it is the fact that no provision is made for M and N to override or

---

34 We do not wish, at this stage, to speculate on "mechanisms" or structural/functional substrates for the various forms of "memory". We do not think that it is feasible to discuss memory in isolation, simply as an abstract set of "storage" and "retrieval" processes. This would, in particular, not provide us with the means to distinguish the memory processes of an organism from those of a computer. Rather, we think that "memory" must be discussed as part of an integrative and comprehensive structural/functional architecture. We believe that certain of the components of the "conceptual architecture" which we discuss in the present paper (especially in Chapter 7), such as "symbolization", "categorization", and "the value system" will be relevant to such a discussion of memory. However, at the same time, we think that this architecture needs to be fleshed out further before one enters on such a discussion. Similarly, as discussed in the second footnote to the introduction to Section 6B, our "module" terminology is as yet too crude and too general to be of real use here as a means of representing "memory" processes.

For the same reasons we shall not, in the present paper, propose specific mechanisms for the various forms of "learning", but shall limit ourselves to some general remarks in Section 6.10 on connections between "learning" and "control". In a subsequent paper we shall experiment with approaches for incorporating memory and learning more explicitly as part of our system architecture.

circumvent the "symbolic" control procedure and to act on E directly. To take an extreme example, this may be necessary in case of "emergency". More generally, the control architecture should allow the system more latitude in "choosing" between direct and "symbolic" control, depending on the circumstances. In other words, the relation between controller and controlled modules is not quite satisfactory in the above picture. We shall have more to say about this in the next section.

6.10 Learning as control and learning to be controlled

a) Learning as control[35]

It may be appropriate to view the learning of "associations" between sensory inputs and actions (or between sensory inputs themselves) as itself constituting a basic form of "control", perhaps still falling under the rubric of control as "gluing of local logics". That is, depending on the modular decomposition under consideration, the formation of these associations may be realized via the alteration of the interfaces between modules (including, perhaps, the modification of the modules themselves, or the incorporation of new modules)[36]. In any case, such learning is a basic concomitant of other control processes. Among its roles are: helping to build up the symbolic representations or "world pictures" discussed in the preceding section, and "facilitating" certain actions of various modules. This form of "control" very likely operates on a slower time-scale than the "interface relations control" discussed above, and may perhaps be viewed as providing a backdrop for the operation of the latter.[37]

b) The learning of control

We have emphasized the picture of control as "interaction among modules" (including the action by one module or class of modules to modify the interactions among others). As with any action of one module on another (or on the interfaces between other modules), this kind of action can be learned (in somatic time) via association and

---

35 We shall be speaking rather crudely here of associations involving "sensory inputs". A better picture of what we have in mind may be obtained from Section 7D on "categorization".

36 The obvious analogy here is "synapse formation" and modification of "synaptic strengths".

37 Very crudely speaking, the former control the nature of the communication channels, and the latter control the "traffic" along these channels.

Even if (as suggested in the text) this form of "control" is realized via the alteration of module interfaces, this still leaves open the question of what kinds of processes carry out, or bring about, this alteration. As in the case of "memory" (cf. the footnote to Section 6.9c)), it is unlikely to be helpful to attempt to "localize" the processes of associative learning (even in a functional vs. structural sense) to a particular "module". More likely, whole networks of modules will be involved. At this stage, we do not wish to speculate as to particular mechanisms; however, we expect, at a minimum, the following conjunction of factors:

(i) Synchrony (or approximate synchrony) of occurrence of certain "inputs" and "outputs". The temporal order in which "events" occur may also be expected to be a factor.

(ii) Biochemical/molecular biological mechanisms for reinforcing/refreshing synaptic connections on a long-term basis. This may include mechanisms of "long term potentiation".

(iii) An internal value system, either primary or higher-order (as discussed in earlier chapters), which is the ultimate arbiter of control.

"reinforcement" processes.[38] In connection with this learning we envision a fundamental role for the internal value system which, in a sense, is the ultimate reference for control (cf. the final footnote to the preceding paragraph).

c) Learning to be controlled

However, this is not the end of the story. As we have seen in Section 6.9c), the relation between controller and controlled module *cannot* simply be one-way, from controller to controlled. This consideration is extremely important, because it may even impact on the survival of the system. Thus, there should be an additional kind of learning: namely, the module should learn both how to be controlled and how to circumvent control when necessary. The module should learn how to reroute its output to a higher-order "planning center" and await permission to act. In some sense, the module should learn patience, yet also irresponsibility (although, as usual, we would not wish to phrase this as a strict and universal rule). Speaking somewhat metaphorically, if certain driving inputs to the module are rather low, the preference for being controlled should win out, but if these driving inputs are sufficiently strong, decisions should be taken at the lower level. Again, this should be an emergent behavior of the module (or of the module/controller combination): the way the module makes use of control should be a learned process.

---

38 In speaking of "reinforcement" we do not mean to limit consideration to behaviorist learning paradigms.

7. Neural Symbolization

In this chapter we shall seek to bring out what are the central characteristics of "symbols" as we conceive the term, and to begin the development of a framework for discussing and analyzing the processes of "symbolization", i.e., the construction and manipulation of symbols. We have already briefly examined "symbolization" in Sections 5B and 6B. As discussed there, we believe that symbolization has a central role to play in the functioning of adaptive autonomous systems. An additional aim of the present chapter is to propose elements of an integrative functional architecture schema centered upon processes of symbolization.

The term "symbol" has many connotations, only certain of which we wish to draw upon, and others of which we wish to de-emphasize. In particular the term will not be used in its traditional linguistic sense, nor as it is commonly used in AI. Thus, as will become clear, our own approach runs basically transversal to the much debated "physical symbol system" vs. "connectionism" polarity. Our thinking on "symbolization" is primarily influenced by, and directed towards, biological organisms endowed with nervous systems; hence, our terminology of "neural symbols". However, in the present paper we are not attempting to scrupulously mirror the neurobiology of specific model organisms, but are only aiming at the more modest goal of biological "plausibility" (in the sense of Section 3.8c)).

A. Characteristics of neural symbols

7.1 Physical description and role of neural symbols

a) Physical description

In a nervous system communications between neurons (and also other cells), and between "modules" is supported by a range of neural symbols. At a small-scale intercellular level these may be realized, for example, as action potentials, neurotransmitters, neuropeptides, or hormones. Within an individual cell, the physical realization may, for example, be via ions, enzymes, and gene regulation. On a larger scale the physical realization may take the form of firing patterns (or, more generally, "activity" pattern, including detailed biochemical states, etc.) of a population of neurons; moreover, this population need not be localized within a small delimited area, but may be (widely) distributed within the nervous system (either "diffusely" or in "clusters"). In particular, a neural symbol should not be thought of as a simple discrete unit, nor as an element of a fixed discrete "alphabet".

b) Role of neural symbols

The roles of neural symbols are: 1) to maintain an ongoing interaction of the organism with its environment both at the levels of perception and action, and involving internal as well as peripheral sensors and effectors. 2) to maintain the integrity and identity of the organism as a system. For all organisms this includes the regulation of the homeostatic support processes. In the case of higher organisms it may also include the constructing and maintaining of some sense of self. These roles are carried out, first, by

the creation of neural symbols at the level of sensor organs and by the action of neural symbols on the effector organs and, second, via interaction between different neural symbols.

7.2 Comparison and contrast with the notion of "symbol" in AI and linguistics

Our concern here is with the notion of "neural symbol", rather than with the notion of "symbol" more generally. Accordingly, we shall not attempt to enter into an in-depth discussion of either AI or linguistics, but shall limit ourselves to some brief remarks, consisting largely of pointers to later sections of this chapter.

a) Comparison with AI

From the perspective of AI, the "system" is a computational engine. Correspondingly, "symbols" are abstract entities, with form essentially decoupled from content, introduced to serve as the substrate for formal computational manipulation by the system. This viewpoint, by focusing exclusively, or predominantly, on the processes of computation, leaves out of consideration the crucial (and more difficult to analyze) non-computational "background" processes which form the essential substrate for the computation. These processes, ultimately carried out externally, by the programmer, include the introduction and "interpretation" of the symbols, i.e., the association to them of content or reference. By contrast, "neural symbols" emerge from within the system, and their content is linked to their physical action within the system (cf. Sections 2.9, 7.4a), and 7.4b)). More generally, as emphasized throughout this paper, the modes of functioning of adaptive autonomous systems are primarily non-algorithmic. Accordingly, the processes involving "symbolization" are primarily not computational in character.

b) Comparison with linguistics

One parallel between "neural symbolization" and linguistic construction is the absence of "type" restrictions (cf. Section 7.3a). In natural languages, statements can refer to any construction of the language and "reason" over it. One can begin, for example, with a purely descriptive proposition, and then treat it as an object, qualify it by predicates, etc.... (Cf. the discussion in Section 7.10 of symbolization by a subsystem, in particular the constructions associated with "higher order" planning).

A basic difference between neural symbols and linguistic expressions centers on "decomposition" into parts. Linguistic expressions are composed of smaller parts, and can be syntactically analyzed. On the other hand, a neural symbol is not a priori analyzable in "smaller parts", but is homogeneous (cf. Section 7.3(c)(v)). Moreover, its semantics is not a composition of the semantics of the syntactic constituents (cf. Section 7.4c)).

7.3 "Homogeneity" of neural symbols

a) "Type" questions in formal language systems, logics, and natural languages

Our use of the term "homogeneous" is nonstandard. We do not seek to suggest by this terminology that neural symbols are internally undifferentiated, nor that they are indiscriminately interchangeable, nor of a uniform complexity. Neither do we seek to

deny the patent heterogeneity of cell types or circuit architectures. Rather, we wish to address the issue of presence or absence of "type" in the formal language sense.

The introduction of "types" is ubiquitous in the settings of formal languages, logics, set theory, and languages for computer science. This is true also of formal language approaches to natural language semantics, such as Montague's intensional logic. The term refers to the imposition of a kind of hierarchy of "levels" among the symbols or expressions in the theory, together with rules of well-formedness for compound expressions based on the "types" of the corresponding constituent expressions.

One notable example of an untyped formal language is the untyped λ–calculus of Church and Kleene. This calculus, originally intended as a instantiation and formalization of the notion of "effective computability", lies at the foundation of the Scott-Strachey approach to "denotational semantics" of programming languages. Very roughly speaking, all the expressions in the calculus are "functions", and functions can act on, or be applied to other functions (including themselves) indiscriminately. It is something of this character, though surely not the calculus itself, which we seek to carry over to the context of neural symbols.

This character is also manifested in natural languages. In a very basic sense the usual language of human discourse is not "typed", because anything can be referred to at any level. Certainly, there are syntactical categories and grammar rules (or principles) to be respected. Still, on the other hand, abstract nouns can be the subjects of concrete verbs, and propositions can themselves be referred to by other propositions, and can be assigned a grammatical function. Also, certain words can have purely context dependent meanings or referents (e.g., pronouns, demonstratives, adverbs, indexicals ...) rather than intrinsic signification. Phenomena such as these are hallmarks of an almost total absence of "type" in a classical logical sense (and, indeed, go beyond the bounds of even a loosened type discipline which admits of "polymorphic" types).

b) A trade-off between paradoxes

"Types" were introduced into mathematical logic by Bertrand Russell[1], and soon thereafter incorporated into Whitehead and Russell's "Principia Mathematica". The reason for their introduction was to circumvent the so-called "set theoretic paradoxes" arising in attempts to formalize the intuitive notion of "set", as well as a related family of "semantic paradoxes" associated with issues of self-reference. An example of the former class of paradox is provided by "the set of all sets which are not members of themselves". Is this set a member of itself or not ? As easily checked, either case leads to a contradiction. (This is sometimes expressed more picturesquely as the "barber paradox". In a town there is a (male) barber who shaves those and only those men who do not shave themselves. Does he shave himself ?) An example of the semantic paradoxes is the "liar paradox": Epimenides the Cretan says that all Cretans are invariable liars. Should we believe him ? A related introduction of a "levels" hierarchy is due to Tarski.[2] He argues, roughly speaking, that to avoid inconsistencies in the use of a "truth" predicate requires

1 See Russell's "Mathematical logic as based on the theory of types", 1908.

2 See A. Tarski "The concept of truth in formalized languages", (translation of 1933 paper).

an infinite hierarchy of meta-languages, each containing the truth predicate of its predecessor.[3]

While (perhaps) circumventing the set theoretic and semantic paradoxes, the introduction of type or level hierarchies seems to bring us up against a worse (we would argue) class of paradoxes, the well known "infinite regress" or "homunculus" paradoxes which are prone to arise in theories of brain organization. A related set of problems centers on issues of "combinatorial explosion" or "computational complexity". Our view, which we shall argue for below, is that a proper schema of "neural symbols" can avoid both classes of paradoxes. (See, for example, our discussion in part c) below, our discussion of higher order planning in Section 7.10c) below, and our discussion in Section 7.12 of avoiding regress to infinity).

c) The case of neural symbols

The homogeneity of neural symbols manifests itself in several different senses, among them the following:

(i) Complexity

The logical "degree" and the content realized by a given symbol are not necessarily reflected in a corresponding complexity of this symbol nor in the symbol manipulation in which it participates. For example, the construction of an association among various symbols is, in general, realized by a neural symbol which is not more complex than the "constituents" of this association. This example (and others to follow) illustrates a basic point: the logical complexity of an external observer's description of the functioning of a neural symbol may inhere in the observer's descriptive framework, and need not be mirrored by a corresponding complexity in the actual neural symbol manipulation.

(ii) The question of hierarchy

A related point concerns the question of "hierarchy". As a case in point, we ourselves, for purposes of description and analysis will be led to speak (as we have already done) of "higher level" or "more abstract" symbols and the like. Thus, it may be convenient to speak of the "type" or "function" of a symbol in order to emphasize certain distinctions. For example, one symbol "emitted" by a "module"[4] may pass directly to an

---

[3] The set-theoretic and semantic paradoxes are still issues of controversy. A good introduction is provided by R.L. Martin "Recent essays on truth and the liar paradox", Clarendon Press, 1984. Another interesting treatment is given by R. C. Koons "Paradoxes of belief and strategic rationality", Cambridge University Press, 1992. Koons argues that it is possible to construct a "type-free computationalism" which nonetheless can circumvent the liar paradox. Moreover he argues that such issues will be crucial in making the transition from rational agent models of the social sciences to frameworks involving institutions, rules, and practices.

[4] We remind the reader that we use the term "module" (discussed more fully in Section 6.2) in a variety of senses. Depending on the setting, we may have in mind "behavioral", "functional", or "structural" modules, associated to distinct (and surely not isomorphic) modes of "subsystem decomposition". Thus, a given module associated with one such decomposition need have no "counterpart" in another decomposition. In particular, the structural (or functional) substrates "supporting" a given behavioral module need not be sharply localized, but may be (widely) distributed. Moreover, this supporting substrate need not be static, but may be history and context dependent.

effector organ or motoneuron, while another symbol emitted by the same module may instead contact a higher order system or module in the "hierarchy".

Yet, just as in the above discussion of "complexity", this "hierarchy" is, at the least, very heterarchical. A given element or subsystem may participate in a multiplicity of organizational modes (possibly concurrently) , with its "rank" differing from mode to mode[5]. Moreover, even within a single "mode" (e.g., some form or instance of planning or decision-making), the attribution of a "rank" ordering may be very much viewpoint dependent. For example, it is misleading to think in terms of rigidly ordered "pathways" for "flow of information", or in terms of rigid "chains of command", with successively higher-order decision-makers sending "information" up the chain and "commands" down the chain. As an example of the subtleties involved, we recall Section 6.10, where we saw that "lower" centers can learn to be controlled. Thus, questions such as "who decides to do planning" are surely badly-posed. Such a formulation ignores facts like the following: The system plans in a "parallel" fashion, and must also learn to plan. "Lower" centers may work (quasi-)independently unless they "fail" (for example, fail to meet expectations) or if they encounter conflicts, in which cases they may request (or require) the intervention of "higher" order centers for arbitration or planning; this process may itself be learned. (The "criteria" for failure may be those of the "lower" center itself or, alternatively, may derive from "higher" centers. For example, the activity of the "lower" center may be linked to the execution, or attempted execution, of a higher-order plan; cf. the discussion of "error monitoring" and "expectation" in the last footnote to Section 7.9d), and also the discussion of "plan revision" at the end of Section 4.4b).) Moreover, in case of "emergency", the hierarchy can be bypassed, with the lower centers overriding the higher centers, and taking control of the whole system.

This absence of an absolute "hierarchy" is due, in part, to the "homogeneity" of neural symbols (in the various senses discussed in this section). Reciprocally, it gives rise to an additional sense in which neural symbols are "homogeneous".

(iii) Specialization

While there is a striking degree of specialization and specificity within neural systems (e.g., as regards connectivity, morphology, intrinsic membrane properties, and transmitter "profiles" of neurons or neuronal subpopulations), this does not necessarily impose a corresponding specialization or differentiation on the neural symbols supported by these cells (or cell populations).[6] To illustrate, suppose in the preceding example (discussed above in (ii)) that the module is a neuron whose axon has multiple collaterals (possibly forming "biochemically" heterogeneous classes of synapses), which project to distinct targets in the nervous system, at varying functional "levels". In this case (unlike in the interpretation above), when the neuron generates action potentials it may be more appropriate not to attach any "type" distinctions to the neural symbols emitted by these

---

5 Even in considering an artificial vs. natural system, it may be appropriate to construct a range of functional and structural diagrams (e.g., consider the design of a large aircraft). There is no reason to expect these all to be simply progressive refinements (i.e., incorporating progressively greater degrees of detail) of a single top-level "master" diagram.

6 At an intracellular level, a related kind of homogeneity, or lack of sharp "type-specificity" of interaction, is mediated via the general mechanism of "allostery", whereby arbitrary metabolic circuits can interact without the need for any direct steric matching of constituents.

different collaterals at the various synapses. As another example where "type" distinctions may be out of place, a firing pattern in one neuronal population may induce firing patterns in a range of other subpopulations, including itself.[7] (As in Chapter 5, we are not concerned in the present chapter with assigning a precise meaning to the term "neuronal firing pattern".)

(iv) Specificity

Much of the specificity of a neural symbol, especially in its details, is strongly dependent on context and on the (internal) state of the system. For example, the "same" neural symbol will never be instantiated twice via exactly the same neuronal firing pattern. Moreover, in contrast with most "mathematical" contexts, this is not a sharply-specified "non"-specificity. Thus, it is most likely inappropriate to speak of an "equivalence relation" in the formal sense (since "transitivity" probably fails to hold). Nor, for that matter, is it clear that the notion "statistical ensemble of firing patterns" is adequate. This "non"-specificity is all the more striking when one considers that it is "iterated"; i.e., neural symbols interact with other neural symbols, giving rise to yet other neural symbols.

A particular manifestation of this context dependence is a kind of "just-in-time" specificity, whereby the full details of the construction of a given neural symbol are not "decided" until the symbol is actually needed to act.

(v) Internal structure

"Homogeneity", in this sense, refers not to an absence of internal structure, but rather to the absence of an a priori, canonical, decomposition of a neural symbol into "smaller parts". There may be no "meaningful" decomposition or, as may more often be the case, several concurrent decompositions, depending on the context. (See also Section 7.4c) below, on non-compositionality of semantics).

7.4 The content of a neural symbol

a) "Acting is being"

We have already mentioned several times that for an embedded system there is no natural content-free syntax, while the opposite is true for a formal or simulated system. This is in particular the case in the setting of neural symbols. Simply by virtue of its mere being, a neural symbol acts, and one might say that the rationale of its being is its action.[8]

---

7 At first glance, the remarks above as to homogeneity of symbols hold also for computers or robots, at the level of their actual hardware. By "symbol", in this setting, we refer to the electrical or magnetic activity patterns sustained within the hardware. However, we draw attention to three basic differences with the natural system context. First, there is the basic role which value systems play in the latter setting (see Section 5B). Second, in the case of computer or robotic systems the primary description is the high-level description provided by the designer or programmer, and its instantiation on any particular machine is simply one realization of this high-level representation. In the neural context, on the other hand, the primary representation is generated by the value system, in conjunction with the creation of neural symbols at the sensor level (because the system is autonomous in the strong sense). Third, in the case of a computer or robotic system the various evaluation criteria for the symbol language are all "external", rather than for the "benefit" of the system itself.

8 Corresponding comments can be made for the symbols used in the hardware of computers, robots, and the like, but with the same caveats as in the preceding footnote.

b) Distinction between syntax and semantics for certain neural symbols

In the present context an insistence on the traditional syntax/semantics distinction is problematical, and may even be inappropriate and irrelevant.

Still, for purposes of analysis, it is possible, at least at the level of sensors and effectors, to attach a semantic content in the traditional sense to a neural symbol created by a given sensory organ, or acting on a given effector organ.

For example, consider a stimulus a and a stimulus b to which the animal is sensitive. (We mean here that there are certain sensory neurons which fire when the animal is exposed to a, and that a neural symbol "a" is generated.[9] Similarly for b. Then the symbols "a" and "b" do whatever they have to do.) Now, it may be of importance for the system to be able to recognize the co-occurrence of a together with b, while a and b alone are of no importance. Then the system will learn to develop an association symbol "ab" which is "emitted" (at a certain higher level) when the animal is exposed to a and b. Thus, for example, "ab" can be a signal for food, while "a" and "b" separately are not associated to food.

A semantic content can also be attributed to neural symbols manipulated by certain higher order neural structures whose functions, in the traditional sense of the term, have been clearly identified. Among the examples which arise are those involving "memory" or, as we shall see below, "planning" modules.

c) Non-compositionality of the semantics of neural symbols

We see, in this example, the main phenomena of the non-compositionality of semantics.[10] The content of "ab" is not a function of the content of "a" and "b". It may not be interesting or useful for the system to be able to reconstruct the symbols "a" and "b" once "ab" has been created. The association symbol "ab" acquires its own content and connotation, triggers other modules (while "a" and "b" would not), makes its own associations (for example the system may associate "d e f" with "ab" but not with "a"), may become a higher-order value (i.e., become tied to the higher-order value system), while "a" itself has no value, etc. A related possibility is that "a" has a certain content (say, communicating with certain modules or inducing certain actions), while "ab" (by virtue of the circuit architecture) will induce unrelated actions.

All this says that the symbol "ab" acquires a new status by itself which is not a function of the content of "a" and "b" separately.[11]

---

9 A word as to our notation. We distinguish the neural symbol "a" from the actual stimulus a via the use of quotation marks.

10 This non-compositionality stands in contrast to the Montague semantics and denotational semantics referred to in Section 7.3a).

11 Remark concerning the logical failure problem: The fact that neural symbols have a function which is their semantics (although for description purposes we may introduce an artificial distinction) is extremely important because it explains why the system cannot enter into a logical failure (see Chapter 4.B). In fact such a neural system does not compute or manipulate symbols in a Turing machine sense (which, as we have seen can induce logical failure) It is clear, though, that the system can fail, since it eventually dies. But this is not the same as logical failure. When a neural symbol is emitted, it actually does actions on the modules to which it is sent and these actions will finally induce certain actions on the environment. On the other hand, the symbols of the tape of a Turing machine do nothing.

B. Encoding

7.5 The combinatorial strategy

a) The basic conflict : richness of the environment / scarcity of resources

An adaptive autonomous system satisfies the principle of relative adaptation: it must maintain a state of relative adaptation with its environment.

Let us recall briefly the issue we face here: in general, the environment is complex and changes. Certain characteristics of the environment can change very easily, but others are rather stable and fixed. On the other hand, the system "extracts information"[12] from the environment (and from its own internal state, see Section 6A) and can act on the environment (so that because of the constraints, the internal state will change accordingly). The problem is that the system has very few sensors (compared to the phenomenological richness and unpredictability of the environment) and it has also very few effectors (compared to the tasks it must do) and moreover the constraints give rise to conflicts (either of sensory type or of action type). But in any case, the basic conflict any adaptive system must face is the conflict between the richness of the environment and the scarceness of sensory and action resources.[13]

b) The combinatorial strategy

The strategy used by adaptive systems to resolve these basic conflicts is combinatorial, both at the sensory and at the action levels.

(i) At the sensory level

A priori, the system does not know what are the relevant associations of sensory data, nor the relevant emergent "categories" which are needed to maintain a state of relative adaptation with its environment, which may vary significantly. Consequently, the system requires a basis for constructing a very broad range of possible complexes of sensory data.

One of main tools available to the system to construct such complexes is the chronological coincidence and/or the spatial co-occurrence of "feature data" and, also, the temporal succession and/or spatial juxtaposition of data in various "modules" of the system.

On the other hand, at least in the case of higher organisms, these complexes of data, by their very construction, will very likely be used by the system in the construction

---

12 We use this terminology merely as a convenient mode of expression, and do not intend thereby to suggest a reification of "information", as something pre-existent in the environment which is simply to be filtered and transferred from the outside to the inside of the system. Rather, we regard categorization of "feature data" to be an active, constructive process carried out by (and upon) the system. We shall discuss categorization at length in Chapter 7D.

13 We do not wish to oversimplify here. The sensors and effectors may themselves be complex, and the interfacing with the environment may be at multiple "levels of abstraction"; cf. the discussion of "layered hierarchical systems" in Section 2.6e), and also the discussion of "sensitivity" vs. "sensors" in Section 3.3c).

of its own chronology [14], and in the elaboration of internal representations of (itself as embedded in) its spatial environment.

(ii) At the action level

At the action level, the system must construct and coordinate a rather large and not a priori specified set of actions; but it has very few effector devices, and is also submitted to constraints which even further reduce the combinatoric possibilities of action. So the system conceives complexes of actions, carried out either in parallel or in temporal succession. A priori, when the system is learning (and it is always, to some extent, learning) it must allow for a broad range of possible combinations of elementary actions. (For the notion of "elementary" action, see Section 5.1b).)

Surely, in the course of the learning process, only a circumscribed number of combinations, either of sensory data, or of elementary actions will be retained, but a priori, the system must allow for many more. (Surely also, evolution has restricted the number of possible combinations by imposing a basic circuitry).

We have spoken above of "sensory" and "action" levels. We do not mean to suggest by these expressions that there is a sharp segregation between these two, either in a functional or structural sense. Certainly, there is communication between the two "levels" via intermediary processes. Somewhat more directly, the system needs to maintain some form of sensory monitoring of the state of its effectors and, reciprocally, the effectors are needed to act upon the physical apparatus mediating the sensory functions. In addition, the sensory-motor functionalities are by no means limited to the periphery of the system (where it interfaces with the external environment), but will be present internally in various forms. Moreover, as noted in Section 3.3c), the sensory apparatus (in conjunction with the effector apparatus) need not function in a purely "passive" manner, but may actively "explore" for "relevant" signals. What the system treats as "relevant" may be context-dependent, and will be linked to other aspects of the system's functional architecture such as categorization, memory, planning, and the value system.

c) Expansion-compression / encoding-decoding

The issue we address now is indicated in the schematic "input-output" diagram shown in Figure 7.1, with data "flowing in" at the sensory side, and action "flowing out" at the effector side.[15] The sensory side of this diagram indicates an expansion of sensory data, whereby the system builds new neural symbols encoding combinations of sensory features (occurring either simultaneously or in temporal sequences). These combinations reflect certain statistical regularities of the environment, as made use of by the organism. There is a corresponding bottleneck on the effector side of the diagram. This is intended to indicate a compression of effector data, whereby the system, starting from certain

---

14 Presumably, the construction of its chronology will also draw upon the presence of "biological clocks", both at a (sub)cellular level and at a more global level.. (As regards the latter, Damasio (1994), suggests such a role for rhythms such as the heart-beat.) A partial list of the "problems" of chronology facing the organism includes the following: the sense of time of occurrence of events, the notion of ordering of events, object "permanence" (i.e., the construction of "objects" on the basis of a temporal succession of snapshots), and "permanence" of the self.

15 This picture of "information/action flow" is not intended to be taken literally. See the footnote accompanying Section 7.5a).

neural symbols encoding combinations of effector features (again occurring simultaneously or in temporal sequences) must decode them to transfer them as actual complexes of actions for execution by the effector system.

The intermediate portion of the diagram indicates neural symbol construction and manipulation by the system. These processes of symbol construction and manipulation by the system are used to "glue" together the sensory and effector sides of the diagram (or, alternately phrased, to "coordinate traffic" between the two bottlenecks)[16]. Near the sensory side (resp., the effector side) of the diagram, neural symbols constructed and manipulated by the system have a semantic content of an essentially sensory (resp., effector) character. As one moves further from the periphery (either sensory or effector periphery) the corresponding semantics of these symbols is deprived of a clear-cut sensory or effector character. This notion of "semantics" can refer either to the organism's internal semantics (as discussed in Section 7.4) or, equally well, to the viewpoint of an external observer.

d) Combinatorial ≠ discrete

In emphasizing a "combinatorial" strategy, we do not wish to emphasize a discrete-mathematics character. In particular, the encoding-decoding processes must be carried out in an environment which imposes on the organism a continual flux of inputs, while requiring from the organism a continual flux of outputs. Thus, the organism is not confronted by an orderly succession of distinct "problems", but by an ongoing asynchronous flux of "problem-complexes". Time (and timing) constraints (see Section 4.1) are a major consideration. As an example of the type of resource allocation involved, the organism must "buffer" varieties of inputs and intermediate outputs until it can allocate the resources necessary to further "process" them.

7.6 The necessity of encoding by symbolization

The question is how to implement this combinatorial strategy in an effective way. It is quite difficult to manipulate complex combinations of features *as such*, or to refer to them, store them, or manipulate them, *as such* . It is clear that, for many purposes, it will be preferable for the system to create a single neural symbol "ab" for the co-occurrence of a and b rather than maintaining and dealing with a pair of distinct symbols "a" and "b" (keeping in mind that these symbols must be physically instantiated). A fortiori, this kind of argument will be valid for the extremely complex combinations of neural symbols that the system will construct, in particular given the fact that these neural symbols may themselves be higher order constructs.

In other words, the system avoids an extensional expression of complexes of "features" (or of other symbols) by encoding such complexes using just a single symbol. This need not imply that the symbols at the source of a given complex are not individually maintained or reconstructable. The fact that complexes have semantics which may be largely independent of the semantics of each individual symbol (as we discussed in Section 7.4 c)) means that each new complex thus created will have a "life"

16 There are fundamental issues associated with this "gluing": What constitutes a "coherent" gluing ? What are the processes or mechanisms whereby this coherence is brought about and maintained ?

of its own. That is, it will participate in other associations as, and be manipulated as, a unit. This is why the system will tend to develop structures which can represent each complex of symbols as a unit.

7.7 Example: a schema for associations

a) There are many aspects to the gluing processes discussed in Section 7.5. Among them, one can distinguish the processes for association. By this we do not simply mean purely sensory-sensory or purely motor-motor associations, but also sensory-motor associations, not necessarily at the periphery. Here we wish to address these issues only en passant, briefly introducing some schemata which we plan to elaborate elsewhere (incorporated into a more comprehensive architecture).

We do not envision a canonical schema for associations. Among the various distinctions that can properly be made, we limit ourselves here to one broad distinction: that between prewired circuitry, on the one hand, and associations which must be learned or acquired in somatic time, on the other hand.

b) Prewired schemata

We have in mind here not only reflex-arcs, or reflex-like associations, but also more complex prewired behavior patterns (such as the "fixed action patterns" of invertebrates).

Even in this setting of prewired schemata there are problems of coherence of behaviors. One is the "singleness of action"[17] which is necessary for coordination of behaviors. Another is the need for task-dependent and phase-dependent modulation of reflexes.[18]

c) Associations requiring learning

Such associations can take place at any level of combinatoric expansion or compression. Nor does "learned" necessarily mean that the association takes place only between highly complex sensory and effector sequences. For example, one can envision the learning of reflex-like associations, i.e., linking to a basic sensory feature a certain elementary action, bypassing the whole architecture of combinatoric expansion and compression.

---

17 This term was introduced by Sherrington [Sherrington, C. , "The integrative action of the nervous system", Yale University Press, 1906] in his studies of the interactions between simple spinal reflexes in mammals. He concluded that "singleness of action" results from a competition between the different reflexes. Similar themes, involving the "behavioral hierarchy" of the mollusc *Pleurobranchaea* have been pursued by Davis, Mpitsos, and collaborators. (An early representative article is W. J. Davis et al., American Zoologist 14 (1974), pp. 1037-1050. A reassessment of the whole issue of rigidly characterized (hierarchies of) behaviors vs. context-dependence of action, etc., is discussed in G. J. Mpitsos and S. Soinila "In search of a unified theory of biological organization: what does the motor system of a sea slug tell us about human motor integration?", pp. 225-290 in K.M. Newell and D.M. Corcos, eds., "Variability and motor control", Human Kinetics Publishers (1993).)

18 This is reviewed on pp. 282-284 of K.G. Pearson, "Common principles of motor control in vertebrates and invertebrates", Annu. Rev. Neurosci. 1993. 16:265-97.

In order to lend a bit more concreteness to this discussion, we will briefly sketch one plausible[19] association schema, which we shall elaborate in a later paper. This association schema involves the following mechanism:

It consists of an association "module" which, a priori (before learning is begun) associates various sensor symbols, value symbols, and effector symbols which are actually put into execution on the environment. When the "correct" effector symbol has been selected and put into execution (and this may be discovered by chance at the beginning of the learning period), it is reentered into the association module. This leads, due to the presence of constraints, to the production of a value symbol. This value symbol, in conjunction with some form of "LTP" (long term potentiation) mechanism linked to the underlying circuitry, gives rise to an association whereby the sensor symbol (which was just previously produced in temporal coincidence with the action symbol which was actually executed) "reinforces" the production of this action symbol.

If no value symbol is produced (because the action symbol put into execution is not the "correct" one), no reinforcement takes place. The notion of "correctness" used here is not of an "instructive" character. In other words, the learning of this association is done without a "teacher", in an autonomous fashion. The basis for this whole process is the fact the organism is embedded in its environment. The corresponding constraints both mediate and provide the rationale for the activity of the value system. Thus, the "correct" action in the environment induces, reciprocally, on the organism the production of a value symbol (which serves as a "reward", i.e., which has a positive valence). This is a consequence of the constraints which induce (or strongly bias) the production of this value symbol. We note here that these values need not be primary, but may be higher order (see Section 5.8). In particular, the value system itself may be only in part hard-wired, and may involve a learning or acquisition process. In a similar vein, we comment that the above association schema is completely general, in the sense that it works equally well at any "higher" level (with sensor symbols being replaced by inputs of certain "modules", and effector symbols being replaced by outputs of certain other "modules").

C. Maintaining autonomy

7.8 Preservation of autonomy and relative adaptation with the environment

a) In Section 3.1 we saw that a cardinal characteristic of natural systems is to preserve their existence and autonomy, while maintaining a state of relative adaptation with their environment. This takes place in the presence of various constraints. A range of these constraints was discussed in Section 4.1, namely homeostatic constraints, effector and sensory constraints, chronological constraints, and constraints on the internal structure of the system. Finally, the resources that the environment affords and that the system is able to draw from it are, in general, limited.

b) The previous facts manifest themselves in several forms. Among these we can distinguish (and this list is not exhaustive):

---

19 By "plausible" we do not mean that this schema bears close similarity to any actual biologically instantiated association circuitry. We mean only that it is not obviously at variance with any biochemical or biological constraints or "principles".

(i) The need to acquire resources, as signaled by the value system. Here we have in mind not only the acquisition of "positive" resources, but also the avoidance of "negative" resources, such as noxious substances or dangerous situations. Either case implies a tendency, or even a necessity, for action.

(ii) Reciprocally, a tendency to preserve its own resources as much as possible. In contrast with the above cases, this implies a tendency to defer actions in the environment.

These two reciprocal tendencies (acquisition of resources and preservation of resources) give rise to the strategy of symbolization, as we shall elaborate later.

(iii) The system tries to maintain an autonomy of choice of actions. This involves the ability to construct a broad range of relevant choices for action while, at the same time, not being overwhelmed by this breadth of choice. That is, the system must maintain the ability to select among the choices constructed so as to act in a timely fashion. This may require the resolution of conflicts not only among competing choices of action, but among competing processes of evaluation of choices.

(iv) On the other hand, in many situations, e.g., if the environment is simple, or familiar, so that "failure" does not occur (cf. Section 7.3c)(ii)), or if no conflicts arise, or in case of "emergency" (requiring reflex-like actions), the system will tend to act in a more-or-less reactive form, without referring to higher order modules. An extreme case is the reflex arc, where the link from sensor to effector organs is direct. This strategy of avoidance of intervention by higher order "centers" realizes an economy of metabolic resources and/or of time resources (as in the case of emergency).

So, roughly speaking, the strategy of the system is to act in a purely reactive way whenever possible, without referring to higher order centers. (However, this statement should not be interpreted too simplistically; cf. Section 7.3c)(ii)). We give an additional illustration: Certain sequences of actions may initially need to be acquired via attentive learning, involving higher centers; however, once acquired, their control and execution can be transferred to an essentially reactive level, bypassing the higher centers.

In many situations, such a purely reactive strategy is too limited, however. First, the organism, in general, needs to perform complicated actions in a rich environment, while subject to the basic constraints at the sensory and action levels, as noted earlier. Second, the environment and the internal state (e.g., metabolic levels , neuronal activity patterns, etc.) can give rise to conflicts which must be resolved.

In either case, both when a reactive policy is adequate and when it is inappropriate, the strategy of symbolization is needed. In fact, certain reactive procedures may not be as direct as a simple reflex arc, and may involve memory processes for "storage", recall, and recognition of relevant constellations of features or patterns of action. Furthermore, in certain circumstances the choice between a purely reactive policy and a recourse to higher order centers, if both of them are possible, itself involves decision-making by the organism. (See also Section 7.3b)).

c) Adaptation is not universal

It is clear also that an adaptive system can adapt and behave intelligently only in certain classes of environments. If the sensory modalities are not adapted to the environment, the living system will probably develop a clumsy behavior which we can qualify as "unintelligent". But we qualify it "unintelligent" only because we know the answer to the decision problem, and we know what we ourselves would do if we were in this situation, and how we could solve the problem. But we don't know what we would do, given the information, the possibilities, and the values that the organism actually has.

Somehow the trivial mistake of (one traditional form of) AI is that it wants to define a system which could adapt to any environment, say a robot able to wash dishes, walk on the moon, have social relations, attend a math class...

d) An example : the bee and the window

Let us consider a familiar example of a bee (say), which is prisoner in a room where a window is still half-open. The bee begins to wander within the room and, because it is attracted by the sunlight, it goes towards the window, where it bumps against the glass. Surely, the bee is not equipped with a sensory modality detecting glass windows at distance (nor, for that matter, are we), so the bee begins bumping and bumping again into the glass. Now, the bee can detect the window where it bumps into it. But the bee, being unable to detect the window at a distance, cannot form a global view of its local landscape (namely, recognize the window is half-open). After a while, the bee stops bumping and usually goes back in the room, until finally the sunlight stimulus attracts it to the window where the bumping behavior starts again. Finally, after a certain number of trials bumping and wandering, the bee, by mere luck, finds the opening and leaves the room. Now the point is not, as some AI practitioners would do, to qualify the bee's behavior as unintelligent. In fact, probably the most clever thing the creature can do, given its sensory possibilities, is precisely what it does: namely, try by force (but not too much , and not too stubbornly) to get out, and then wander for a while to try to find a better opportunity. Thus, in some sense, the bee's "strategy" is the best one.[20]

7.9 The general symbolization strategy

a) In general, because of the complexity of the environment and the complexity of the actions to be performed, the system will "try" to develop (on an evolutionary time scale) or to use (on a somatic time scale) structures which circumvent the need to continually carry out repetitions of the same kind of analysis and synthesis of sensory data or of actions. This leads inevitably to the construction and use of memory mechanisms ("storage", recall, and recognition) in conjunction with processes of categorization. This implies that complexes of sensory features or of actions must be encoded in an effective way so as to be easily manipulated, "stored", and "retrieved", whenever necessary. When a new situation is encountered, and when this situation is recognized as "similar" to a situation previously met and dealt with, the system need not engage in a lengthy and costly process of random or exhaustive search to select or construct an appropriate course of action. Instead, drawing on its processes of memory

20 This viewpoint may sound somewhat Panglossian, yet be correct for all that.

and categorization, together with its value system, it need only consider *relevant* possibilities for action, which are further modulated and refined by the specifics of the new situation.

b) Memory

At present, aside from the association schema sketched in Section 7.7, we have little to say on the complex and controversial issues surrounding "memory" in its various forms.[21] However, we do wish to make some brief remarks on memory in connection with neural symbols.

(i) Short term memory

Short term memory is indispensable for building neural symbols registering the combinatorial temporal succession of short sequences of basic sensory features or of elementary actions (or of other, more abstract, neural symbols).

A precise circuitry for building such neural symbols will be described in a later publication. The question is, essentially, to "stop" time for a short period. As we shall show, this can be realized (not so easily) by reentrance and "short term potentiation".

(ii) Long term memory

Long term memory will be used to build an efficient planning schema (so that the organism can draw upon previous experience and not always have to plan the same course of actions). We have, as yet, no biologically plausible[22] scenario for building such a memory circuitry.

c) Planning and decision-making

In case conflicts arise, it can be advantageous to the system not to actually venture new courses of action, which may fail to be successful, or may even be disastrous. This is the reason why the system (at least in the case of higher organisms) develops and uses planning structures (see also Section 6.9). Planning allows the system to avoid engaging immediately in real actions on the environment.

Instead, it carries out: (i) Virtual actions on internal representations of the environment; (ii) Virtual evaluations of the results of these virtual actions. These actions and evaluations, while virtual with respect to the external environment, are carried out by actual manipulations, via neural symbols, of the internal representations. (iii) Once this process has been carried out, a "correct" virtual action is chosen and "translated" into a neural symbol for the execution of an action on the external environment, which is

---

21 The literature on learning and memory, from invertebrates to humans, is vast, ranging over a multitude of experimental, clinical, theoretical, and computational approaches drawn from many disciplines, and encompassing work at all levels: molecular, cellular, "systems", and intact (or brain-damaged) organism. In particular, in the settings of neuropsychology and cognitive (neuro)science, various distinctions have been drawn, e.g., "procedural" vs. "declarative", "episodic" vs. "semantic", "long term" vs. "short term" vs. "working" memory. In our brief remarks in the text we shall make a broad distinction between "long term" and "short term" memory. However, at least at present, we use this terminology only in an intuitive, rather than technical sense.

22 In the sense discussed in the footnote accompanying Section 7.7c). Surely, a long term memory anything like a computer RAM fails, on many counts, to meet this plausibility requirement.

eventually carried out. (iv) This virtual action may be associated to the picture of the environment, and this association may be stored for future use.[23]

We wish to stress that the process of virtual evaluation, as well as the concomitant decision-making process, have to be learned, and are carried out using the value system (see Section 5B). We also emphasize the fact that this evaluation process is not simply an optimization procedure in the usual sense. More basically, it is not an "effective", i.e., algorithmic, procedure (although it may "call upon" algorithmic procedures).

The external/internal dichotomy introduced above can usefully be refined as follows: one part of the internal environment may be external relative to another, and corresponding processes of planning and decision-making arise in this setting. For example, the process of deciding between a purely reactive behavior (if available) and a more complicated planning process is of the same type as the decision process discussed above. Thus, this discussion carries over to the present setting. We shall further elaborate on this in Section 7.10 below.

d) Persistence of attitude and shift of attention

Typically, a behavior is not punctate in time, but involves a sequence of action complexes. Thus, a related class of problems that we shall need to address is how the system architecture enables the organism to strike the proper balance between "persistence of attitude" (i.e., having begun a given course of action, following it to its intended completion) and "shift of attention". By the latter we mean, roughly speaking, the interruption of the current course of action if it appears likely to fail, or becomes inappropriate, or if another goal acquires greater "salience".

This type of question is well illustrated by the example discussed above (see Section 7.8d)) of the bee and the window. The organism should persist in its attitude, but not insist too much. If it has no persistence of attitudes, it's behavior becomes completely incoherent and erratic and can be qualified as "unintelligent", not because we decide so, but because the system does not maintain adaptation even in a rich environment. For example, suppose an animal is at a certain distance from two pieces of food (and it sees or smells the food at distance). One could imagine that the animal[24] would hesitate and approach one piece of food, then the other one, doing a kind of oscillatory motion between the two and finally starving in front of two pieces of food. That this fact does not occur is due to the persistence of attitude; namely, after perhaps a short period of hesitation, the animal decides to choose one piece of food, and is not troubled by the presence of the other piece of food.

On the other hand, this persistence of attitude must not be pushed too far. As an example, we can consider an animal separated by a small obstacle from a piece of food. The animal will bump against the obstacle (trying to reach the food) but its attitude (namely following some gradient line) should not persist too much, and the animal should finally go around the obstacle, or perhaps wander around trying to find another source of food.

---

23 We will not address here the basic question of how the passage from actual to virtual, and vice versa, is effected and learned by the system.

24 Traditionally, Buridan's ass.

Achieving the desideratum that attitudes not be stubbornly persistent when they reach no success may involve various "levels" of the system's circuitry. For some purposes, as in the example above, reactive processes may suffice. One such process might be a natural habituation or fatiguing of the firing of the "module" controlling forward movement, and a sensitization of the sensory module registering the bumping against the obstacle. Still, the animal must associate (maybe in a permanent form) the "move forward" neural symbol[25] with the sensory neural symbol elicited by the bumping. At a higher level, issues of "error monitoring" and "expectation" may be involved.[26]

7.10 Symbolization by a subsystem

a) Adaptive subsystems of an adaptive system

We have seen above that the basic tendencies of an adaptive autonomous system lead to the development and use of symbolization procedures. Assume, for the sake of argument, that one has identified a set of modules of the given system which can, in some sense, be regarded as a subsystem. It is useful, for various purposes, to consider this subsystem as having an environment of its own. This environment is that "portion" of the world with which the subsystem interacts directly. This includes the other subsystems of the whole system and, perhaps, a "portion" of the environment of the whole system. We shall refer to this environment of the subsystem as its "proper" environment.

The subsystem receives inputs from its proper environment and acts by outputs upon it. These inputs are neural symbols created by other subsystems, or may be sensory inputs from the environment of the whole system. The outputs of the subsystem are neural symbols which are created in other subsystems as a result of their interaction with the given subsystem, or may be certain effector outputs on the environment of the whole system. As in the case of sensors at the periphery, i.e., interfacing with the external environment, the subsystem need not be simply a passive recipient of inputs from its proper environment, but may actively "explore" for "relevant" signals. (The remarks in the footnote to Section 7.5a) also apply here.)

In any case, the subsystem has its own processes of neural symbolization, and will make use of such symbolization for the same reasons as discussed above for the whole system. At least, this will be the case provided that this subsystem is sufficiently adaptive, and presents a sufficient degree of autonomy relative to its proper environment. In particular, it will tend to develop and use memory devices, planning devices, etc., relative

---

25 This labeling of the neural symbol as "move forward" is not intended to be construed in an overly simplistic manner.

26 As discussed by Swash [M. Swash, "Order and disorder in the motor system", pp. 113-122 in C. Kennard and M. Swash, eds., "Hierarchies in Neurology", Springer, 1989] this monitoring for errors may involve not only input from the periphery, but also an "internally perceived feeling of a mistake". He notes that monitoring strictly from the periphery could introduce unacceptable delays, and consequent cumulative errors via negative feedback. He thus suggests that such peripheral monitoring may be more relevant in conjunction with the learning of a "motor program" than for guiding the execution of a skilled motor performance.

The role of "expectation" in motor organization is discussed by J. Kien, in the setting of hierarchically-nested time frames, where each time scale provides a long term context influencing the lower level time scales. See J. Kien, "Remembering and planning: a neuronal network model for the selection of behavior and its development for use in human language", pp. 229-258 in K. Haefner, ed., "Evolution of information processing systems", Springer, 1992.

to itself. In addition, it tends to develop means of pattern categorization and recognition, for patterns produced by its proper environment.

b) External localization

These "devices" for memory, planning, etc., may be physically localized within the subsystem itself (as in the case where the "subsystem" is the whole system). But in addition, the subsystem may draw on portions of its proper environment as an auxiliary mode of localization. Similarly, "internal" representations constructed by the subsystem may actually be localized outside the system, in its proper environment. That is, while the "logical" localization may be internal to the subsystem, the "physical" localization may be external to the subsystem. This will turn out to be significant (see Section 7.12 below) in avoiding "homunculus" problems. (Such "external localization" is, in effect, a basic concomitant of Damasio's "convergence zone" schema for memory [cf. Section 2.18a) and the footnote to Section 7.12(iii)]. In this setting it also serves to circumvent "homunculus" problems.)

c) Example: "higher order" planning

In connection with the following discussion, the reader may refer to Figure 7.2. We consider a system S with a certain "basic level" set of modules L which are fairly directly connected to the external environment of S (through the sensor and effector devices). The system S develops a set of planning modules P which construct internal representations of the environment of S, on which P carries out actions corresponding to virtual actions of S on the environment (as we have seen in Section 7.9c)). Now let us consider P as a subsystem of S. The whole system S will "try" to develop a set of "higher order" planning modules M which play the same role in relation to the actions of P on L that P plays in relation to the actions of S on the external environment. This set M will construct internal representations of the "environment" of P (constituted by L), on which M carries out actions corresponding to virtual actions of P on L.

Stated more fully, M "plans" in the following sense:

(i) It constructs internal representations of the inputs from L to P. These inputs would be neural symbols "signaling" the possible requirements of L for intervention and eventual control by P.

(ii) It builds neural symbols acting on these internal representations. These actions correspond to (or may be regarded as) virtual actions of P on L.

This means precisely that M "controls" the decision-making procedure "deciding", in effect, whether L can act directly on the external environment, i.e., without requiring the intervention of the planning modules P, or if, instead, L will act on the external environment only after having required the intervention of P. In some sense, M is "thinking" or "reflecting" on the representations[27] that P constructs of L, and on the possibility of actions of P on L. Succinctly, M plans the planning by P.

---

27 A word of caution as to terminology. The "representations that P constructs of L" are not to be confused with the "internal representations" that P forms of the external environment of S. They form,

This is already quite complicated, but even more is necessary. The organism has, in fact, to learn to carry out this whole process. Moreover, P also has interfaces with other subsystems as well as L and M, a notable case being its interface with the value system. In addition, at least in principle, the above construction can be carried out repeatedly, at successively "higher" levels.

7.11 Decision-making procedures

We have briefly alluded to decision-making in connection with "planning", in our discussion of the general strategy of symbolization. Here we wish to comment on decision-making in the wider sense.

a) Artificial systems

For unembedded systems (such as an AI program), the designer typically attempts (in principle) to anticipate in advance[28] all the "situations" that could possibly arise, and specifies in advance, if not the specific decisions to be made by the system, at least the decision-making "policies" that the system must use. Generally, this specification takes the form of "If ... then ..." behavior rules which "fire" via a process of "pattern-matching" against the "state" of the environment. The state of the environment is explicitly characterized in "propositional" terms, and updated by the system via the application of another set of rules specifying how a behavior undertaken by the system alters the environmental state. In addition the designer specifies the "search" policies to be used by the system to generate pattern-matches.[29] A somewhat less rigid form of behavior and decision-making specification, allowing for a kind of "adaptation" by the system is provided by the ("genetic algorithm" based) "classifier systems" of Holland.[30]

A prototype example of how "behavioral choice" is handled in the context of embedded systems ("autonomous" robots) is the "subsumption architecture" of Brooks. (We shall treat this as a specific architecture, but it is, more properly speaking, an evolving set of principles or guidelines for architecture design). Here the designer builds (via hard-wiring, which includes the use of a formal programming language, i.e., the "behavior language") a set of behavior-generating modules, together with an essentially hard-wired priority hierarchy among the modules. The over-all behavior of the robot is purely "reactive", with switching between behaviors mediated "through the world". That is, the behavior of the currently active module brings about a state-of-affairs (in the real world) which triggers the action of a different module (which, at the same time, inhibits the action of the original module). The hierarchy need not be completely rigid or linear. Thus, some degree of context-dependence of the prioritization is possible. However, this

---

rather, the analogue of the representations that S forms of the external environment, via the neural symbols constructed by its sensors.

28 This does not imply that the designer necessarily "gets it right" the first time, or the n-th time. There is a lengthy design-test-redesign-test... loop.

29 The specifications for state change can also be at a meta-level, i.e., specifying changes in the system's rules or search policies, etc.

30 See, for example, J. Holland, R. Nisbett, and P. Thagard, "Induction: processes of inference, learning, and discovery", MIT Press (1986).

context-dependence of behavior is itself hard-wired and reactive, and does not involve "choice" by the robot.

It is important to note here an important contrast between these embedded robots and the unembedded AI programs discussed above. As we saw in Section 5B, it is essentially impossible to give a formal-language specification corresponding precisely to one of our natural language characterizations of a "behavior", e.g., "avoid obstacles". Thus, the real world "behaviors" which result from the hard-wiring are only approximations (and this was all they were ever expected to be) to the "informal" or "vague" natural language "behavior specifications" which are taken as the points of departure for the design of the hard-wiring. This process of approximation to "vague specifications" can help mitigate the "brittleness" resulting from a purely formal approach but, at the same time, can introduce new problems. Thus, as progressively more modules are incorporated, a purely reactive mode of behavioral coordination can give rise to "incoherence" of behavior. For one thing, even assuming that some form of (context-dependent) priority hierarchy is adequate, how can the designer work out in advance what it should be ? Second, it is clear that "emergency" situations can arise where low level modules should seize control. It is unlikely that even a context-dependent hierarchy can accommodate this type of policy shift. It seems that some modes of autonomous (as opposed to pre-wired) decision-making are necessary.[31]

b) Methods of decision - making

It is clear that the decision-making issue is of primordial importance. This is a problem which is solved all the time by any living system however primitive.

There are(at least) three possible modes of choice:

(i) prewired choices
(ii) situated choices
(iii) planning.

The mode of prewired choices is essentially the one used for artificial systems (as discussed above). In the context of biological systems, it is probably reserved for choices or decision-making which have survival value, but that the system would not be able to learn in somatic time. However, prewiring, of itself, does not permit adaptation or the development of "intelligence".

c) Situated choices

Here, depending on the situation of the environment, the system may depart from the strategy of prewired choices, and learn new associations of sensory inputs and effector outputs. At the minimum, this requires time. The problem now arises: What is the (instantiation or realization of) the decision procedure which decides when to shift from prewired choices to learned choices ?

---

31 An early version of the subsumption architecture is presented in R. A. Brooks, "A robust layered control system for a mobile robot", IEEE Journal of Robotics and Automation, Vol. 2, Number 1, pp. 14-23, 1986. A discussion of the incoherence problem, phrased as a problem of "lack of modularity", is presented (together with a proposal for a solution, via control interfaces among the modules, incorporating "motivation" and "arbitration") in R. Hartley and F. Pipitone, "Experiments with the subsumption architecture", Proceedings of 1991 IEEE Conference on Robotics and Automation, Sacramento, CA, April 1991, pp. 1652-1659.

d) Planning

We have already discussed this in Sections 7.9 and 7.10. Again, the question arises of knowing how one decides to change from situated or learned choice to planning.

So, it seems that one gets a regress to infinity, the decision-making centering now on the choicebetween choice-modes.

7.12 Avoiding a regress to infinity in decision-making

The discussion above of decision-making and planning (see Sections 7.9 - 7.11) raises the spectre of an infinite regress in decision making, requiring the intervention of ever higher order decision-makers. The potential sources of trouble include: the choice between choice policies (see Section 7.11); successively "higher order" planning (see Section (see Section 7.10); and the possibility of infinite "oscillation" or "dithering" (see the discussion of "persistence of attitude" in Section 7.9d)). We shall give several arguments below as to why, in the kinds of system architecture we are considering, such problems of regress to infinity are apparent rather than real.

(i) Finite termination

The most straightforward observation is that there is no a priori reason why the system should, in any actual instance, need to avail itself of more than a small, in particular finite, number of "levels" in the above sense. (Indeed, the second-order level of "higher order" planning, i.e., the level of the modules M discussed in Section 7.10c), may well be sufficient).

(ii) Homogeneity of neural symbols

Another escape from this trap is provided by the homogeneity of the neural symbol system, as discussed in Section 7.3c). (See, in particular, paragraphs (i) and (ii), discussing the viewpoint-dependent character of "complexity" and "hierarchy" in the neural system setting).

(iii) Internal representations can be external

The succession of "functional" internalizations associated with successively higher "levels" need not imply corresponding physical internalizations of the actual circuitry supporting these successive "levels". These realizations may, in fact, share common supporting circuitry. [Cf. Section 7.10b) above, discussing "external localization"].[32]

---

32 This argument is reminiscent of the manner in which A. Damasio circumvents homunculus problems in his proposal for a systems-level architecture for memory and consciousness. (See, for example, A. Damasio, "Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition", Cognition , 33 (1989), pp. 25-62). In this proposal, the integration of various "feature-based" sensory and motor fragments does not require a succession of mappings bringing all the "information" together centrally in one place. Instead, the "information" is brought together in time, i.e., via the synchronous reactivation of the original neuronal firing patterns (corresponding to the original feature images), at their original sites, the primary sensory and motor cortices.

(iv) The value system

We believe that the value system (cf. Sections 5B) will play a fundamental role in preventing infinite search, or infinite "dithering", or the construction of an actual "infinite" sequence of control. (See also Section 7.9a). It is worth reiterating here that the value system is a concomitant of the actual embeddedness (or embodiment, in the case of an organism with a "self") of the system. This holds in a direct sense for the primary value system and, hence, also (though perhaps less directly) for the "higher order" values, which ultimately derive from (and in many instances are functionally linked to) the primary value system. Thus, the value system circumvents the "infinities" by means of a grounding in the world, via the physical embodiment (and "self-involvement") of the system.[33]

(v) Grounding via constraints

Additional modes of grounding the system in the world (contributing to the circumventing of the "infinities") come from the various constraints associated with embedding in the world (see Section 4.1). These, in fact, tend to merge imperceptibly with the primary values discussed above. We illustrate this with some examples.

A case in point concerns the metabolic variables. Even if a system does nothing, its metabolism degrades. This cannot be allowed to persist indefinitely, and so serves as an internal "motivation" for (relevant) action.

---

33 The role of the body (in the sense of an envelope separating the individual from the external environment) as a ground reference for the individual's construction of conceptual categories is argued by M. Johnson and G. Lakoff. (See, for example, M. Johnson, "The body in the mind: the bodily basis of meaning, imagination, and reason", U. of Chicago Press (1987); G. Lakoff, "Women, fire, and dangerous things: what categories reveal about the mind", U. of Chicago Press (1987)). A major emphasis of this work is the proposal to replace traditional "objectivist" semantics by a kind of "cognitive" semantics.

In G. M. Edelman's theory of neuronal group selection a system of evolutionarily "selected" value units, or value repertoires (roughly speaking, associated to basic survival needs) plays a key role in biasing and shaping the organism's behavior, in particular its "perceptual categorization". (See, for example, G.N. Reeke, L.H. Finkel, O. Sporns, and G. M. Edelman, "Synthetic neural modeling: a multilevel approach to the analysis of brain complexity", in G.M. Edelman, W.E. Gall, and W.M. Cowan (eds.), "Signal and sense: local and global order in perceptual maps", J. Wiley & Sons, New York (1989).)

In A. Damasio's "somatic marker" hypothesis (see A. Damasio, "Descartes' error", forthcoming, 1994, Putnam, Grosset Books) a system of (visceral) body-based, emotionally-mediated value markers (both "positive" and "negative"), becomes associated (via the individual's nervous system) to the consequences of various courses of action, on the basis of the individual's experiences. This mechanism of somatic marking plays a fundamental role in guiding and shaping the individual's decision-making. The clinical case studies presented by Damasio, in which medial frontal lobe damage leads to disruption of the somatic marking mechanism, strongly suggest that in the "real world" (i.e., in the domain of "social cognition" vs. formally-structured artificial problem domains) a purely "rational" decision-making process, not grounded by a body-based value system, often leads to infelicitous consequences. Among the mildest of these consequences is a tendency towards "infinite" dithering.

The use of a formal (vs. physically embedded) type of "grounding" to truncate a potentially infinite "interpretation" process is put forward by P. Parikh in the setting of communication between "rational agents" using a shared "situated" language (see "Language and strategic inference", Ph.D. dissertation, Stanford University, Dept. of Engineering-Economics, 1987). Parikh, drawing on Barwise and Perry's situation theory and on game theory, discusses how (under appropriate hypotheses) the structure of the "embedding circumstances" of the communication (as reflected in the structure of the corresponding "ambient game") relieves the agents of the need for the potentially infinite set of nested intentions postulated by Grice's approach to communication.

On the other hand(as mentioned in Section 7.9c), the repetition of the same action without any success for a long time may lead, via metabolic or biochemical constraints, to the fatiguing or habituation of the module issuing the neural symbol for that action, and this module will consequently stop "firing" . This may be contrasted with another mode by which harmful actions may be terminated, namely via the (primary) value system. Thus, if the system acts in a stubborn way, for example bumping and bumping against an obstacle, certain sensory devices may fire so much that "pain" will be induced, so that it will stop bumping.[34]

D. Categorization

7.13 Preliminary remarks: the process of categorization

We begin with an intuitive delimitation of the notion of categorization before, in the next section, giving a characterization of this notion more closely linked to the framework of this paper. Intuitively speaking, "categorization" is the process of classification carried out by the autonomous system (or "organism") in order to build coherent "pictures" of the world, and of the actions of the organism on the world. Here, "world" includes both the environment and, to varying degrees, the organism itself.

a) Categorization as a process

We emphasize "categorization" vs. "categories". Categorization is a process carried out by the organism, on the basis of its neural architecture, and in conjunction with its continual activity in its environment. "Categories" are not pre-existent externally; rather, they result from the process of categorization by the organism. A concomitant of this viewpoint is that a category cannot be treated in isolation, but must be considered as embedded in a categorical architecture which is physically realized by the system's neural architecture. This neural architecture is never static, since the system is in continual interaction with its environment, and is continually learning (with the processes of learning and of categorization bootstrapping off one another). Thus, the whole categorical architecture continually undergoes alteration, and is highly individual and history dependent.

The process of constructing a new category is not simply a process of appending or adding a new entity to a pre-existent collection of other entities, themselves unaltered by this addition. Rather, it in general involves modification of the whole architecture. This process of modification has a steady, on-going character, in contradistinction to the abruptness with which concept acquisition may enter the awareness (in higher organisms).

b) Link of categorization to the value system

The categorization by the organism emerges from the following conjunction of factors. First, the categorization is constrained by the sensory capacities and effector capacities of the organism. Second, a necessity for the categorization process, and

---

34 For a discussion of pain "pathways" (and links to somatic-based value- markers) see the book by Damasio cited in the preceding footnote.

learning processes in general, is the repetition of similar neural symbols, produced by similar experiences. Third, similar experiences are possible precisely because the environment presents regularities of association of data features, due to various constraints. Fourth, the organism is constructed in such a way that its value system produces neural symbols in coincidence with experiences having relevance (either positive or negative) for the organism. In this way the value system attaches valence to various associations, and thereby biases the processes of categorization, and of behavior more generally. (Cf. Section 7.12(iv). As in that section, we refer here both to the primary (or prewired) value system and the higher order value system, introduced, respectively, in Sections 5.6 and 5.8).

In other words, the value system channels the categorization process of the various experiences of the organism into modes "relevant" (either positively or negatively) for the organism. As at various points in this paper, there is a *non-vicious* circularity at work here: What is "relevant" determines the value system and, reciprocally, the value system determines what is "relevant".

c) Link of categorization to behaviors

As we have emphasized in Section 5A, the notion of "behavior" in the context of autonomous systems does not have a sharp-edged character. In particular, one ordinarily cannot give an extensional or operational definition of a given behavior. This should not be surprising. For purposes of "thinking about", planning, and deciding about its behavior, the system must make use of the perceptual/conceptual categories available to it. Hence, one should expect the "vagueness" of its behavioral "specifications" to be congruent to the "vagueness" of its conceptual architecture. [Cf. Sections 4.4b) and 4.4c)]. This (for humans) is mirrored by a corresponding lack of "sharp-edgedness" to natural language (lexicon, syntax, semantics, etc.).

This "vagueness" is a reflection of richness and suppleness, rather than of imperfection and deficiency. True, for an external "observer" who wishes to formalize and/or simulate the system's behavior, this seeming lack of precision may be a source of difficulties. However, for the organism itself, this lack of "specificity" is highly apposite and appropriate.[35] For example, in directing its own activities, the organism need not determine the sharp instantiation (and details of execution) of a particular behavior until the time of action, and these details can be strongly conditioned upon both the external and internal context. (In the literature on (voluntary) motor control is not uncommon to find references to "motor programs". Were one to use such a computer metaphor, one might speak here of a highly context-dependent "compilation" process. However, it is important not to interpret this too literally. True "compilation" involves an algorithmic, top-down, translation from one "high-level" formal language to another "low-level" formal language. The present process is something much subtler, being neither algorithmic, uni-directional (in particular, not top-down), nor a process of translation in any precise sense.)

The "vague" character of the construction of behaviors and categorization processes is consonant with the character of the physical substrate of these processes,

---

35 Putting the matter somewhat simplistically, being susceptible of formalization and simulation is unlikely to confer any selective advantage on an autonomous system. (Here "selective" refers to natural selection vs. selection by a designer).

namely neural symbols and their manipulation. For example, two neural symbols can never be completely identical in a physical sense, i.e., as neuronal activity patterns and detailed biochemical states (see Section 7.3c)(iv)). (This, however, is not meant to suggest that there is any "vagueness" inhering in "individual" neuronal activity patterns or biochemical states). Another, more subtle, consideration is that the production of neural symbols will, in general, modify the architecture of the system, e.g., in connection with learning processes.

d) Types of categorization

For purposes of the present paper, we use the term "categorization" in a broad sense, without seeking to distinguish between distinct "types". For example, we intend this term to encompass both perceptual and motor categorizations.[36] Also we are not concerned here with the dichotomy between (the "lower order") "perceptual" and (the "higher order") "conceptual" categorization, which figures in the approach of Edelman.[37] In any case, such distinctions are somewhat blurred in our setting (see Sections 7.3 and 7.5 dealing, respectively, with the homogeneity of neural symbolism, and the architecture of expansion-compression supporting the combinatorial strategy).

In a related vein, in discussing categorization here we do not insist on concomitant processes of awareness, consciousness, or properly linguistic activities (in the case of humans).

However, we shall seek to determine lines of demarcation between purely reactive systems (which, as we shall see below, do not carry on categorization activities, properly speaking) and genuinely autonomous systems (which can, and perhaps must, support such activities).

7.14 A schema for categorization: modules observing other modules

a) The purpose of categorization

The system, embedded in its complex environment, has to make decisions regarding courses of action, and must resolve the ensuing conflicts among the various decision-making processes. The purpose of categorization, in abstracto, is to assist the system to resolve its conflicts, and to arrive at decisions in an efficient way. (As another example of the *non-vicious* circularity discussed earlier, categorization also plays a role in giving rise to, and helping give form to, such conflicts). We do not intend here to suggest that these decision making procedures, or the categorization associated to them, must be of a conscious character. (In particular, we want to stress once more that categorization is not a priori linguistic, and does not require language as a vehicle. Surely,

---

36 Edelman (see the references in the following footnote) has emphasized the critical role of motor activity as part of the processes of perceptual categorization by the organism. This is consistent with the fact that one "aim" of perceptual categorization is to facilitate actions (relevant actions), on the internal as well as external environment. In speaking here of "motor" categorization we have a somewhat different emphasis in mind, namely the necessity of the organism to categorize its own (potential) motor activities. This categorization can certainly be expected to take place at a "pre-conceptual" as well as at a "conceptual" level..

37 See G.M. Edelman, "Neural Darwinism", Basic Books (1987); "The remembered present", Basic Books (1989).

humans' capacities for categorization are greatly expanded (and modified) as a result of interface with and access to a system of language, but this is a different question.)

b) An example

To make things more concrete or pictorial, we describe a very simple example of categorization. We consider a system which, among many modules, contains a basic set of modules L in direct relation with the environment, so that L is sensitive to stimuli a and b, and produces neural symbols "a" and "b", accordingly. Suppose also that L is able to form the association a+b when a and b are co-occurrent. This means, among many other things, that L produces the neural symbol "ab". Moreover, we assume that L is able to act on the effectors in certain ways.

We refer now to Figure 7.3. In this diagram we have two planning modules P and Q. These planning modules "contain"[38] internal representation modules which "observe" the behavior of L. This means the following. When symbol "a" is produced by L, the module P can "observe" it, which means, in turn, that a neural symbol ""a"" is produced inside the internal representation module of P. We also assume that P can observe "ab" (so that the symbol ""ab"" is produced); but we assume, on the contrary, that P does not observe L sufficiently well, and does not notice the production of "b" (so that no symbol ""b"" is produced).[39] P contains another module producing "symbolic" actions, namely neural symbols acting on the internal representation module of P, and eventually rewarded by the "symbolic" value system of P, which is under the control of the value system of the autonomous system. (In fact, it may be appropriate to include under the heading of "value system", comprising both the primary and the higher order value systems, such "symbolic" value systems as well).[40] These symbolic actions act on the symbolic representations inside P, and when a "successful" action is produced, it is transferred over (by neural symbol manipulation) to an actual action, which is then effected. The above-noted "success" is a symbolic one. The corresponding actual action may or may not result in an actual success (see, for example, the footnote on "error monitoring", Section 7.9c)). We shall return to this point below.

The structure of the planning module Q corresponds to that of P, but with the roles of "a" and "b" reversed. That is, Q observes the production of "b" (but not "a"), and so produces ""b"". Also, it observes the production of "ab", and so produces " "ab"' ". (Notice here the extra ' which is intended to indicate that " "ab"' " is not the "same" as ""ab"" in P. These two higher order symbols are distinct because they are constructed with respect to the distinct modules Q, P (even though this construction may be physically mediated and localized by identical modules)). Like P, the module Q has a symbolic-actions module, acting on the internal representation module of Q, and a

---

38 Not necessarily in the sense of physical localization (cf. Sections 7.10b) and 7.12(iii)).

39 Iterated quotation marks indicate a symbol of a symbol. As emphasized earlier (see Section 7.3c)) in connection with the homogeneity of neural symbols, this logical "type" distinction does not imply a corresponding distinction as regards physical realization.

40 While utility functions play no role in our setting, the notion of symbolic value, as used here, is curiously reminiscent of the notion of "symbolic utility" introduced by R. Nozick. (See p. 48 of "The nature of rationality", Princeton University Press, 1993). According to Nozick, symbolic utility, rather than being a different kind of utility, is a different kind of connection , namely symbolic, to standard types of utility. "The symbolic utility of an action A is determined by A's having symbolic connections to outcomes (and perhaps other actions) that themselves have the standard kind of utility...".

symbolic value system of its own (distinct from that of P), under the control of the value system of the autonomous system. In general, P and Q would work in parallel, and would issue symbolic actions which are different (both at the symbolic level, and at the level of the corresponding actual actions), since these actions are adapted to different symbolic values.

We are now ready to discuss what we mean by "categorization" in this setting. We say that a is categorized by P, but not by Q, because: (i) P "observes" the production of "a" (while Q does not); (ii) P can *symbolically* manipulate the stimulus a; (iii) P can evaluate the result of its symbolic manipulation, using a certain symbolic value system V_P. In the same way, stimulus b is categorized by Q but not by P. Moreover, ab (the co-occurrence) of a and b) is categorized both by P and Q, but in different ways. For example, the symbolic actions that P and Q carry out on their respective internal representations may not be the same, and the symbolic value systems of P and Q, namely V_P and V_Q, may be unrelated.

This raises two questions concerning conflict resolution: (i) How does the system go about choosing between conflicting "recommendations" coming from two distinct categorization/evaluation schemes which have both been incorporated into the system architecture? (ii) What determines which of the incipient categorization/evaluation schemes that emerge are actually retained and incorporated into the system architecture? The first question is not of direct concern to us here. In any case, it is a conflict of "classical" type, suggesting the application of classical solutions: e.g., P and Q may reciprocally inhibit one another or, alternatively, jointly be under the control of another ("higher" or "lower" level) module R, etc.). We shall, however, briefly address the second question.[41] In our view, the "evaluation" of incipient categorization/evaluation schemes is tied to the potential mismatch, noted above, between "symbolic" and actual success. If, over a range of contexts encountered by the system, the extent of mismatch is low, the corresponding categorization scheme will be reinforced or "stabilized"; otherwise, it will be weakened, or allowed to weaken. (Notice that this tacitly requires that some actual actions be performed on the environment, so that a basis for evaluating mismatch exists). There may also be weakening as a result of "competition" with other categorization schemes that may arise, or as a result of alteration in the degree of mismatch (due, for example, to a change of contexts). We do not wish, at the present time, to propose a specific "physical" mechanism for effecting such stabilization or destabilization; however, it will be a mechanism acting at the level of (the physical realization of) neural symbols and symbol manipulation.

c) Towards a definition of categorization

The previous section points to the following general definition of categorization. Suppose we have a set of modules L receiving input from their proper environment (as defined in Section 7.10a)). Denote these inputs as a, b, c, ..., ab, bc, ..., abc, ... (with the juxtapositions indicating co-occurrence). We may also take into account temporal sequences of such inputs. Then we would say that the input a is categorized by an

---

41 It is this second question which is, in effect, central to the neuronal group selection theory of Edelman. (See the footnotes to Sections 7.12.(iv) and 7.13d)). As part of his strong emphasis on selectionist vs. instructionist processes, Edelman is also concerned to provide a general mechanism for generating the diversity on which selection can act.

"observing" module P, if the production of the neural symbol "a" in L induces the production of a neural symbol ""a"" by P, in an internal representation module of P on which P can symbolically act, and for which P has a symbolic value system able to evaluate the results of symbolic actions.

We wish to stress three points: (i) The input a may itself be a neural symbol, since a need only be an input from the proper environment of L, rather than an input from the external environment of the system as a whole. Similarly, the discussion above regarding evaluation of categorization/evaluation schemes can be transposed to the present setting, with "proper environment of L" substituting for "external environment". (ii) The categorization is relative with respect to an "observing" module. Thus, the input a can be categorized differently by various distinct modules P, Q, R, ... each observing L. (iii) As noted several times above, localization with respect to the abstract functional architecture need not correspond to localization in the sense of physical realization. For example, from the standpoint of the functional architecture the internal representation modules of P are kept distinct from the representation modules in the set of modules L. But with respect to physical realization, the two may coincide. (Cf. A. Damasio's retroactivation schema, referenced in Section 7.12(iii), footnote).

In light of point (iii) we can allow for the possibility of two modules L and P observing each other. Moreover, P is a part of the proper environment of L and, as such, creates neural symbols which can be inputs of L. As a consequence, P may be able to categorize its own neural symbol outputs, at least towards L, via the reciprocal inputs it receives from L. Two "special cases" are worth noting: (i) The case when P is the whole system, and L is the external environment. Here we may speak of the system as carrying out a categorization of the environment (and its actions on the environment). (ii) The case when P and L are both taken to be the whole system (or, alternatively, when L is taken to be the whole system plus external environment). Here we may speak of the system as carrying out a categorization of its own internal symbol manipulation as well as of the external environment.

d) Remarks about "sameness" and linguistic categorization

Suppose we start with a set of modules L to which two stimuli a and a' are presented (say, from the external environment), and which issue two neural symbols "a" and "a' " (assume simultaneously, for simplicity). We suppose that the module P is able to categorize each of the two stimuli. We could say that a and a' are categorized as the same by P if the neural symbols ""a"" and ""a' "" produced by P are the same.[42]

---

42 This, of course, raises the question of what it means for two neural symbols to be "the same" (see Section 7.3c)(iv)). There are several intuitive answers, all of the following type : "They are the same if, via interaction with the same neural symbol, they (would) yield the same neural symbol ". This is yet another illustration of the (non-vicious !) circularity endemic to the neural symbol setting. There is a tempting resemblance of this "definition" of sameness to the various notions of equivalence of processes (e.g., "bisimulation equivalence") discussed in theoretical computer science in connection with R. Milner's CCS (calculus of communicating systems) approach to concurrent process theory. (See, e.g., R. Milner, "Communication and concurrency", Prentice Hall, 1989). However, the differences between the neural symbol setting and the CCS setting are profound. For example, the thrust of CCS-like approaches is towards "invariance"; i.e., towards implementation-independence (both hardware and software), context-independence, etc. . In this setting, any particular realization is intended as an instantiation of the same pre-existent abstraction. In the neural symbol setting, on the other hand, everything is highly dependent on the

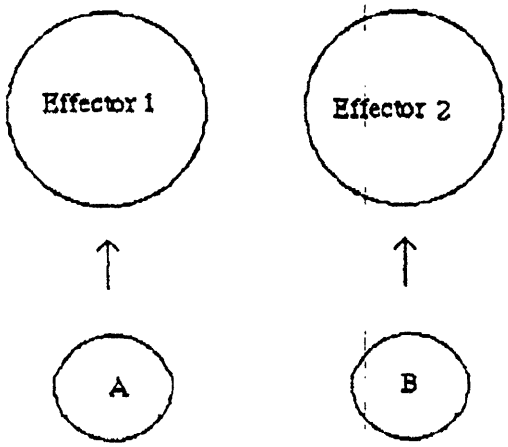
Somehow, P abstracts away the features distinguishing a from a' . If a and a' are objects they cannot occupy the same physical position in external environment, so we can assume that there will be some other module Q observing L which will not categorize a and a' as the same. (Here we will not enter into a discussion of "object permanence", i.e., of how two stimuli a and a' presented at two distinct times could possibly be categorized as "the same" by P. This would lead us, at the minimum, into issues of "memory" that we are not prepared to address at this time.)

From this perspective, linguistic categorization is the categorization done by sets of modules P which, in addition to properly non-linguistic modules, incorporate or interface with very specific linguistic modules which exist only in humans. For example, the ability to use language can modify the over-all categorization architecture, so that even "non-linguistic" stimuli, e.g. visual presentations, might be categorized differently than they otherwise would be. Linguistic judgments as to the "sameness" of a and a' are tantamount to the categorization of a and a' as the same by these modules P. What is special in the case of linguistic categorization is that these judgments can also be communicated linguistically. Moreover, this communication may itself invoke the concept "same", as does the question: "Are a and a' the same ? " This not being a formal language, one can draw a range of responses: not only "Yes" or "No", but "I don't know", "I don't care", ...

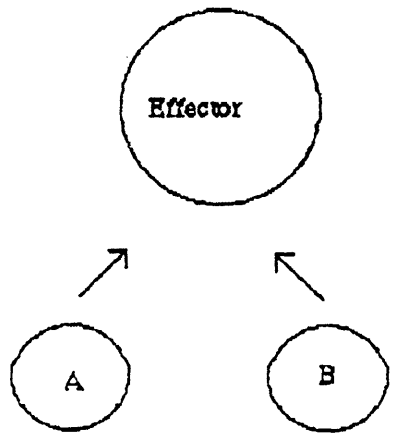
---

individual system, on its history, and on the context. Moreover, the neural symbol system "emerges" from the particular physical realization , rather than being an "instantiation" of anything.

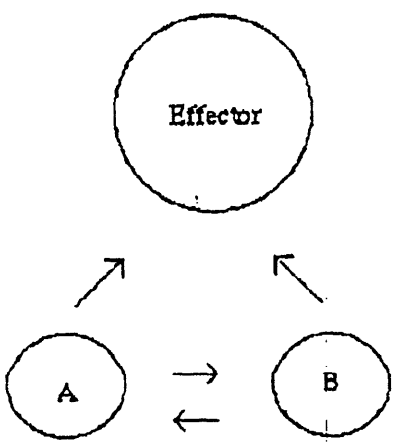
Figures 3.1



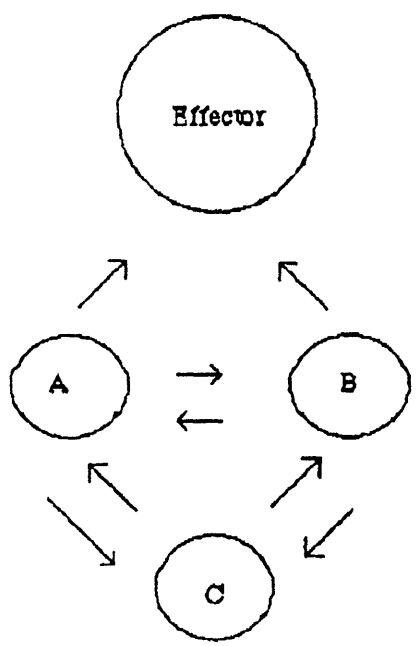
(i) parallel juxtaposition



(ii) juxtaposition : A and B act on the same effector



(iii) A and B act on the same effector and control each other



(iv) A and B act on the same effector, control each other, are controlled by C and report to C .

Figure 7.1

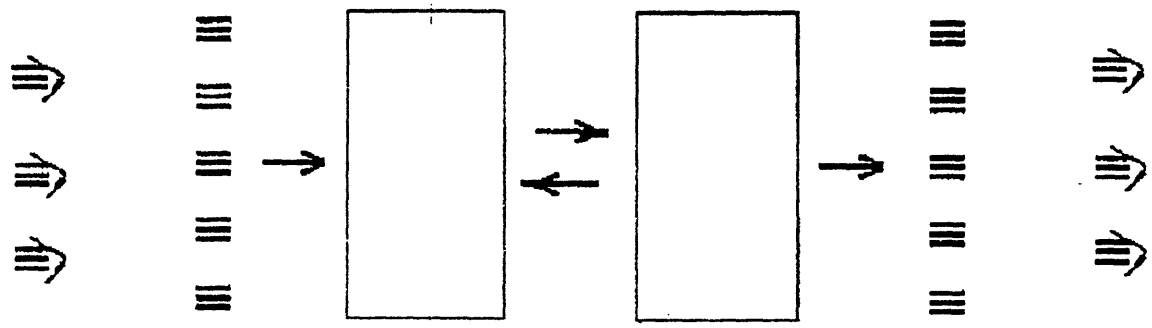


Figure 6.1

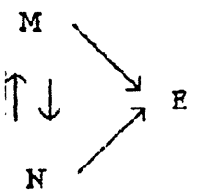
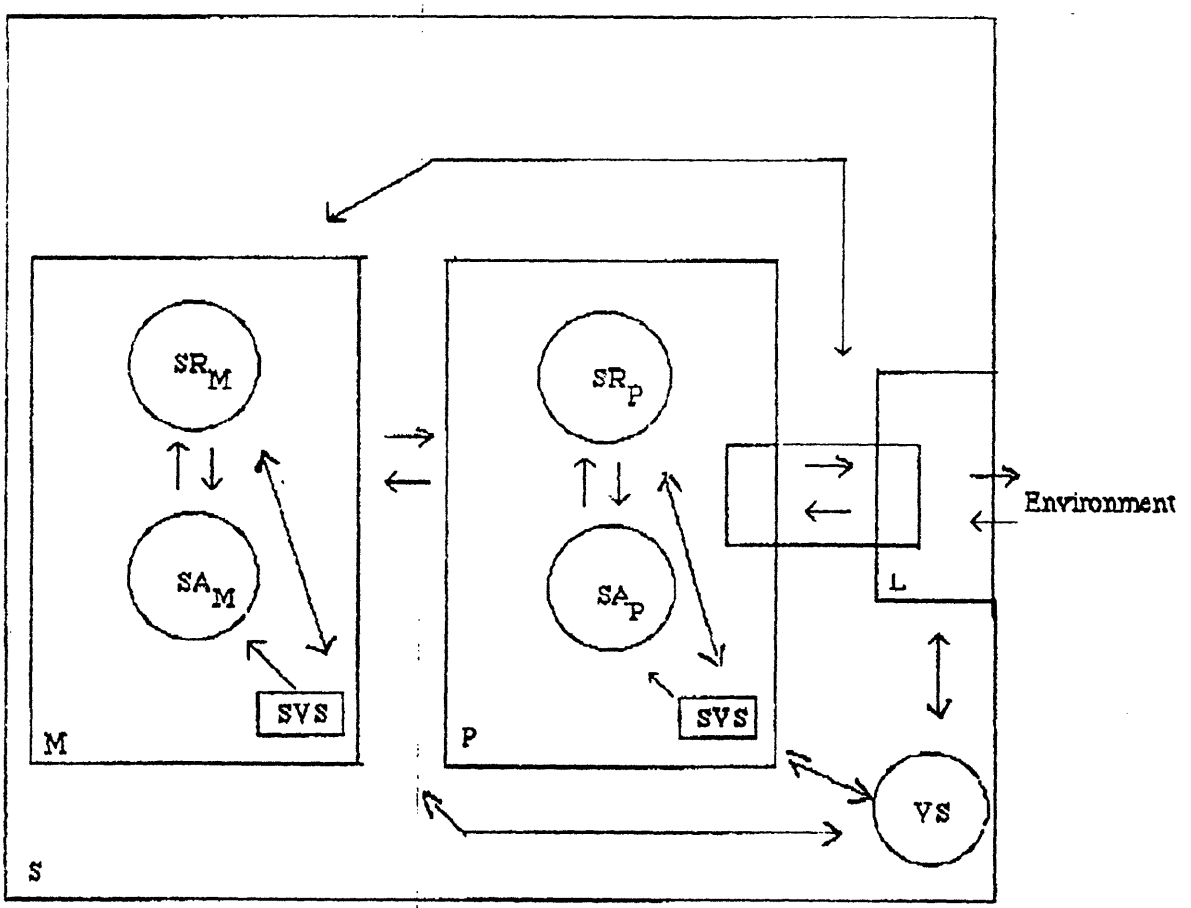


Figure 7.2



planning module

Figure 7.3

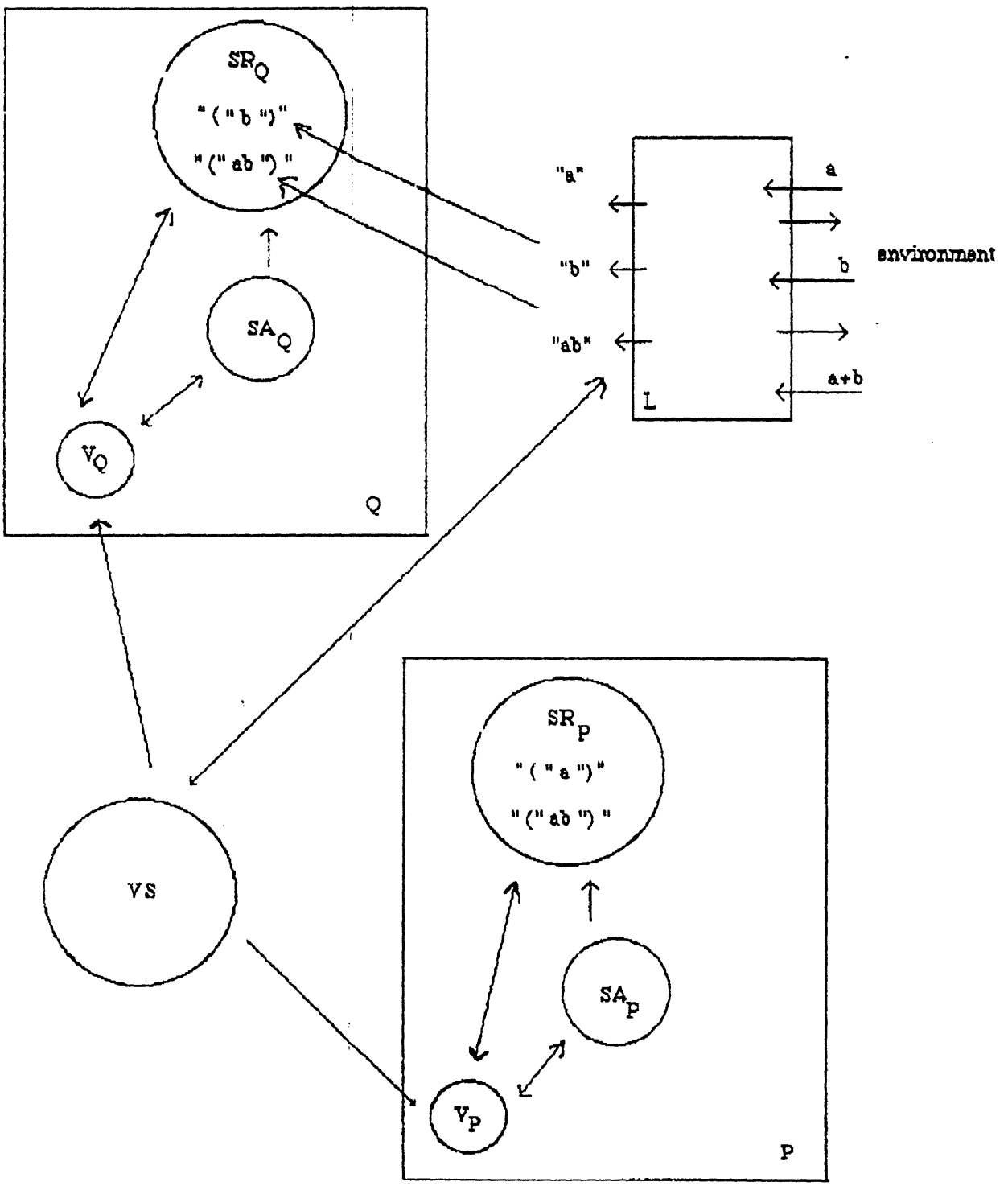


Figure Captions

Figure 7.1

On the left side, a few sensors extract information from the world. This information is treated combinatorially. Various neural symbols representing associations are created. A complex association is encoded by a simple symbol (this is a data compression), but the list of data is expanded. On the right side, very few effectors act on the world. Again, a complex parallel or temporal association or pattern of action can be encoded by a simple neural symbol : this compressed data must be expanded to be expressed by the few effector resources.

Figure 7.2 - Planning module

The system is in contact with its environment through a reactive module **L**. A planning module **P** is constituted by a symbolic representation SRp, symbolic actions SAp (acting on the SRp) and a symbolic value system SVS. This module P makes an internal representation of the environment and can act symbolically and evaluate symbolically the results of its symbolic actions : it makes plans and transfers these plans to **L**. Module **M** observes **P** and the relation between **P** and **L**. It makes an internal representation of **P** and of the relation between **P** and **L**. It contains the same kind of constituents as **P**.

The value system VS evaluates the state of the environment and the results of actions of L on the environment. It also creates symbolic value systems SVS inside P and M for evaluation of their respective symbolic actions on their internal symbolic representations.

Figure 7.3 - categorization

On the right side, the environment contains stimuli a, b,... either in isolation or in combinations a+b... The reactive module creates corresponding neural symbols "a", "b", "ab",... and can act directly on the environment. It can happen that a, say, is categorized in P, because the creations of the neural symbol "a" is observed by P and internally represented in SRp under the form of a neural symbol "("a")". In the figure, a and a+b are categorized by P, but not b. While, b is categorized by Q and a+b is categorized in a different way by Q.