YOU DON'T NEED A SOFT SKIN TO HAVE A WARM HEART
Towards a computational analysis of motives and emotions

Aaron Sloman and Monica Croucher
University of Sussex
Brighton, BN1 9QN
England

Abstract

The paper introduces an interdisciplinary methodology for the study of minds of animals humans and machines, and, by examining some of the pre-requisites for intelligent decision-making, attempts to provide a framework for integrating some of the fragmentary studies to be found in Artificial Intelligence.

The space of possible architectures for intelligent systems is very large. This essay takes steps towards a survey of the space, by examining some environmental and functional constraints, and discussing mechanisms capable of fulfilling them. In particular, we examine a subspace close to the human mind, by illustrating the variety of motives to be expected in a human-like system, and types of processes they can produce in meeting some of the constraints.

This provides a framework for analysing emotions as computational states and processes, and helps to undermine the view that emotions require a special mechanism distinct from cognitive mechanisms. The occurrence of emotions is to be expected in any intelligent robot or organism *able to cope with multiple motives in a complex and unpredictable environment.*

Analysis of familiar emotion concepts (e.g. anger, embarrassment, elation, disgust, pity, etc.) shows that they involve interactions between motives (e.g. wants, dislikes, ambitions, preferences, ideals, etc.) and beliefs (e.g. beliefs about the fulfilment or violation of a motive), which cause processes produced by other motives (e.g. reasoning, planning, execution) to be disturbed, disrupted or modified in various ways (some of them fruitful). This tendency to disturb or modify other activities seems to be characteristic of all emotions. In order fully to understand the nature of emotions, therefore, we need to understand motives and the types of processes they can produce. This in turn requires us to understand the global computational architecture of a mind.

There are several levels of discussion: description of methodology, the beginning of a survey of possible mental architectures, speculations about the architecture of the human mind, analysis of some emotions as products of the architecture, and some implications for philosophy, education and psychotherapy.

Note: Aaron Sloman is now at the University of Birmingham (http://www.cs.bham.ac.uk/~axs/)
Some of the ideas relate to Chapter 6 of *The Computer Revolution in Philosophy* available here:
http://www.cs.bham.ac.uk/research/cogaff/crp

Contents

Introduction
The space of possible mental mechanisms
The nature of our world: constraints on intelligent systems
The computational architecture of a mind
Processing motives
Types of motivators
Decision-making cannot be completely rational
Could a machine have its own motives?
Methodological comment
Emotions: an example
Towards a more general account of emotions
Emotions, moods and attitudes
Could a machine have emotions
Conclusion
Acknowledgements
Bibliography
Appendix

Introduction

Most work in Cognitive Science seems very myopic - concerned with some tiny fragment of a mind, and usually only human minds. Although such detailed analysis is essential, long term progress may require some attention to more global issues: what is the global architecture of a human mind and what alternative sorts of minds are theoretically possible or actually to be found in animals or machines? An attempt should be made to create a generative taxonomy (a 'grammar') of possible types of minds, human, animal, or artificial. In other words, we should attempt to survey the space of possible mental mechanisms, charting the major discontinuities in the space, and their implications. This will provide a context for efforts to explore a tiny corner of the space. The task requires the collaboration of several disciplines.

As a focus for the discussion, we shall attempt to demonstrate the following conjecture:

> An intelligent system, i.e. one which is able to cope with many complex, relatively independent, goals, in a changing and partly unpredictable and uncontrollable world, requires mechanisms whose total effect includes the ability to produce emotional states. So no special emotional subsystem is required for emotions.

Notice that this conjecture says nothing about *single-minded* intelligent systems which pursue a single top-level goal, and only goals subservient to it. Neither does it apply to systems with complete knowledge of the world, able to predict and control the environment without error. Such systems do not need the mechanisms which generate emotions.

What exactly are emotions? The ordinary concept is not very precisely defined, but includes states of being, angry, irritated, elated, embarrassed, awestruck, afraid, miserable, jealous, intrigued, exasperated, impatient, indignant, joyfully expectant, and so on. We all have a lot of *implicit* knowledge about the differences and similarities between these states, for we understand stories, gossip, and the answers to our enquiries about how others feel, and we understand that different emotions can explain different behaviour, thoughts, and so on. But it is not easy to *articulate* this common-sense knowledge. Later we offer an analysis of emotions as states of the total system in which processes generated by some motives tend to interfere with or modify processes generated by others. This provides a framework for analysing and distinguishing ordinary concepts, extending the techniques of conceptual analysis described in Sloman l978 chapter 4.

For instance, the sudden discovery of great impending danger, should interact with the desire to avoid harm, and produce a powerful motive to do something about the danger. This may require considerable disruption of other activities, with rapid re-organisation of thoughts, plans, and possibly physical movements. This will constitute the state, and, if there is internal self-monitoring, the experience, familiar to us in some forms of fright. (Having a certain sort of emotion is not the same thing as experiencing, or feeling it: others may perceive that you have an emotion, such as anger, which you do not feel.) Similarly, disappointment about failing an examination, or excitement about passing, will tend to disturb attempts to concentrate on other activities. Anger involves believing that someone has done something you disapprove of or dislike, as a result of which you wish to harm that person, and thoughts about what has been done and what you might do in return are hard to set aside.

This suggests that at least some emotions are computational states essentially bound up with cognitive states and processes, including perception, beliefs, inferences, goals, planning, deciding and attention. A more detailed analysis is presented later.

The space of possible mental mechanisms

What is intelligence? There is no clear boundary between systems which are and those which are not intelligent. Minimally, we can specify that an intelligent machine must have beliefs and goals. Very crudely: beliefs are internal representations which tend to change when the environment changes, and goals are internal representations which tend to produce changes in the environment. (For a ‘self-conscious’ mechanism the environment will include the internal state of the system.) For now, we assume that the notion of "representation" is well understood.

Using a liberal concept of representation, a thermostat is an example of the simplest sort of mechanism satisfying this minimal notion, at one extreme of the space of possible mechanisms. The internal representation of the actual temperature is a rudimentary belief, and the representation of the "desired" temperature a rudimentary goal. Some very simple life forms are comparable. More complex systems will differ in very many ways. For instance, goal-directed computing systems use more sophisticated modes of representation for their goals and beliefs: structural descriptions rather than a mere scalar variable.

> It is not worth arguing about which mechanisms are "really" intelligent. The important thing is to understand the range of possibilities, and to examine the implications of the presence or absence of various submechanisms, or global architectures.

The space of possible intelligent systems is not a continuum: there are many discontinuities giving it a rich discrete, non-linear, structure. A full survey would need to explore many sorts of discontinuities. A study of different kinds of animals and their abilities (e.g. Lorenz 1977) would help to draw attention to important subspaces. Equally, an overview of different subspaces should help to organise research on animal capacities, and the evolution of intelligence. Some fairly obvious discontinuities include the differences between systems which do and systems which do not have the following abilities:

* The ability to represent not just measures but also structural descriptions in beliefs and goals.

* The ability to represent some goals explicitly, instead of having them all ‘compiled’ in functional mechanisms.

* The ability to represent different goals at the same time: this requires some mechanism for choosing between goals.

* The ability to vary priorities of goals according to circumstances and needs, instead of having a permanently built in hierarchy.

* The ability to integrate different perceptual abilities: e.g. detection of chemical gradients, touch, hearing, sight (2-D vision, 3-D vision, motion perception).

* The ability to record potentially useful information without knowing at the time whether it will be useful.

* The ability to store results of computation to avoid recomputation in future.

* The ability to pursue different goals at the same time, possibly interleaving them, e.g. performing different actions, or solving different problems in parallel.

* The ability to construct plans in advance of executing them.

* The ability to modify plans and other stored information in the light of new information.

* The ability to perceive internal states of the system, and to make use of this self-monitoring in planning, acting and learning.

* The ability to assess and choose between goals in the light of structural matches with rules or the result of inferences (rather than always using some scalar measure or some kind of weighted sum of such measures as the sole basis of choice).

* The use of a collection of more or less independent co-operative subsystems, each with some intelligence, forming a larger more intelligent whole.

* The ability to take advantage of different sorts of 'cognitive friendliness' in the environment (Sloman et al. l98O): e.g. it is possible to infer some regularities in the world on the basis of a sample of observations, and it is possible to use redundancy in the optic array to cope with degraded visual information, or to speed up processing. Not all animals can take advantage of all kinds of cognitive friendliness.

This is just a sample of the sorts of divisions to be explored, in the space of computation systems, natural, artificial, or merely hypothetical. We shall find that a recurring theme, already illustrated here, is the trade off between the flexibility gained by *decompilation* and the speed gained by *compilation.* By decompilation we mean replacing fixed and automatic mechanisms by explicit and alterable rules or descriptions which are 'interpreted' by more general mechanisms. By compilation we mean the reverse process, which apparently occurs during some forms of learning, for example when mental or physical reflex responses are developed. We also include under compilation the process of storing large numbers of specific consequences of general rules or beliefs, and accessing these consequences directly in future instead of re-computing them. There is a tension between compilation and decompilation as the need for flexibility and adaptability vies with the need for speed. The essence of decompilation is that it postpones decision making until the 'run-time' context is available.

An important feature of *human* intelligence is the ability to cope with a complex and partly unpredictable and uncontrollable environment in the light of a large and not necessarily static or consistent collection of motives. It does not seem to be the case that many other animals can do this, and certainly no AI systems do it yet. One of the important philosophical questions is whether it is possible for any artefact to have *its own* goals as opposed to merely serving goals of the designer. We shall suggest a positive answer, provided the machine has enough different sorts of motivation including motivators which modify the set of motivators.

Mental mechanisms cannot be completely ordered in respect of degree or amount of intelligence. System A may be more intelligent than system B in some respects, less in others. For instance, it might be argued that an ant colony forms a system which is in some respects more intelligent than a human being, since more different concurrent processes can be devoted to a common set of goals, though in other respects it is less intelligent, since it does not produce intelligent behaviour derived by reasoning about an explicit set of goals and facts (as far as we know).

A full exploration of the global structure of the space of possible systems would be a major undertaking requiring the co-operation of philosophers, biologists, psychologists, anthropologists and system designers. For now we wish to explore some features of a small sub-space close to human intelligence. Starting from a specification of some constraints which it is plausible to assume influenced the evolution of the human mind, we shall try to derive likely submechanisms. It turns out that mechanisms apparently required for fulfilling the constraints are also capable of producing emotional states. The argument depends essentially on the fact that the unpredictability and uncontrollability of the environment, and the interactions between different motives, requires mechanisms which allow processes produced by some motives to disturb or modify processes produced by others in the light of new information. This is one form of decompilation. Some of the disturbances will constitute emotional states. Later, we indicate

what the mechanisms might be. (See appendix A for a partial overview.) Some of the mechanisms are primarily concerned with achieving goals and ensuring that important goals gain priority over others. Others derive from the need to learn from experience.

There is not yet a computer implementation: at present, our design ideas outstrip current implementation capabilities. In this situation, theories are inevitably speculative, and testing is difficult. Moreover, internal computations may have a richness and complexity incapable of being manifested in behaviour in all their detail. For that reason we may have to be prepared, in sketching designs for possible intelligent systems, to admit specifications of internal processes that cannot be fully tested behaviourally.

It is not clear whether existing computational concepts will provide an adequate framework for the design of explanations of human abilities. It is very likely that we shall discover the need for new ways of thinking about processes. For instance, it may be that the study of "neural-net" architectures will reveal new more powerful concepts and techniques. Certainly, ordinary human processing involves far more parallelism than is to be found in existing Artificial Intelligence models. (Sloman 1981. Schaffer 1981.) However, the level of discussion in this paper is intended to be independent of such issues: there is no commitment to implementations based on present day computer architectures, programming languages, etc. Starting from task analysis, we discuss designs likely to be useful in explaining some global aspects of human intelligence.

In a very interesting paper, apparently ignored by AI researchers, Simon (1967) proposed an analysis of emotions partly similar to our own. He relates emotions to interrupt mechanisms, as we do. However, we shall show that interrupting is but one way in which a motive can affect processes involving other motives, in emotional states. Moreover, we attempt to embed the analysis in a more explicit theory of the processes of deliberation, decision, and action, and we are not so committed to a serial model of mind as Simon. Although this work was begun independently of Simon's it can be seen as a development of his ideas. He identified some, but not all of the processes we shall explore. Moreover, he was primarily concerned with the analysis of human processing, whereas we are ultimately more concerned with the theoretical task of understanding a range of possible mechanisms (cf Sloman 1978, chapter 2 on the aims of science). Some aspects of our analysis are also close to the analysis of emotions presented by Ira Roseman (1979), though we feel that he does not do justice to the complexity and variety of emotional processs.

The nature of our world: constraints on intelligent systems

Our task analysis surveys some of the constraints on mental processes required for deliberation and action on the basis of multiple independent motives, given only partial knowledge, and the possibility of error. Later we use all this apparatus to offer a preliminary analysis of some emotion concepts.

In what follows we frequently refer to motives. The word 'motive' is used here in a very general sense. It covers not only goals, but also 'second order' motives, that is principles for comparing goals, and 'goal generators', which interact with new information to produce new goals. The term 'motivator' will be used sometimes as a reminder that we are talking about both first-order and second-order motives. Examples of motivators are what we commonly call desires, preferences, dislikes, ambitions, principles, ideals. (Later we see that these work in different ways.)

Implicit, or 'compiled' motives are embodied in mechanisms which merely produce certain behaviour. Explicit motives are embodied in representations which can generate processes of deciding and acting, using mechanism whose behaviour may be varied by varying the representations. Some 'higher level' motives may be implicitly embodied in the explicit representation of 'lower level' motives. The less that is compiled into unchangeable mechanisms, and the more that is explicitly represented, the greater the chance that the system can change its motives to cope with new circumstances: an important type of discontinuity in the space of possible mental mechanisms. (This point will be repeatedly illustrated below.) Existing work on planning and problem solving in AI gives some fragmentary models of how motives may be represented explicitly, and how they can generate planning and acting (e.g. Sussman 1975). Notice that the representation of motives need not be different from the representation of factual beliefs. What is different is their role in deliberation, planning, acting and emotions. We describe this role in some detail below.

We shall assume that we are dealing with organisms or artefacts capable of being driven by many different, and independent, explicit motives. Even so this does not uniquely determine the design constraints. Which features of the world are important will depend in part on the type of system. The constraints relevant to immobile animals will be different from those relevant to a species which moves about. Whether the young are mobile or able to defend or feed themselves soon after birth will make a difference to what sorts of computational processes are desirable in both young and adults. Across the whole range of types of animals (and possible robots) there is so much variety, that we cannot hope for a *simple, general* theory of intelligent systems.

However, a systematic theory can be generated by examining types of environmental constraints, then exploring different combinations of relevant information-processing mechanisms and strategies. This defines a taxonomy, or 'generative grammar' for types of minds, analogous to the periodic table of chemical elements, though probably more complex. We should not expect, however, that all the possible sorts of minds generated in this way are to be found in nature. Ultimately, we would hope to be able to relate this sort of task analysis to studies of the evolution of mind and behaviour, though the difficulties are enormous since behaviour and mental processes cannot be fossilized.

Some of the constraints have already begun to be analysed in AI research especially vision research. Our list of constraints concentrates on issues concerned with deciding and acting. We shall see that a recurring theme is the need for different processes to occur in parallel to maximise speed and flexibility, and for some processes to interfere with others. Many of the processes need not be conscious.

We can distinguish constraints due to the nature of the physical environment, constraints due to the nature of the body, constraints due to the nature of the brain, constraints due to being part of a social system, constraints due to the helplessness and under-development of infants, and constraints due to the requirements of mental processes postulated as a result of other constraints, for instance, the need to learn from experience, which derives from the complexibility of the environment, making it hard to get things right first time.

* The need for structural descriptions.
Goals, obstacles, opportunities, food, friends and foes do not come with simple physical patterns to identify them. Moreover, they can be perceived from different viewpoints and in different circumstances. Consequently simple physical sensors, or pattern recognition algorithms will not suffice for intelligent goal-directed systems. Rather, they will need to be able to interpret sensory information by building and comparing structural descriptions. Motives will need to incorporate structural descriptions of states to be achieved, preserved, avoided, or terminated, rather than, for example mere measures. Detection of satisfaction or non-satisfaction of motives will often involve complex perceptual processes of inference and interpretation. These are some of the reasons why concepts of control theory are inadequate for analysing or designing intelligent systems. What kinds of decompilation, referred to above, are possible, depends on the power of the representational apparatus available. Much work in AI has addressed this sort of representational constraint, and we shall not discuss it further. (See also Sloman 1978, chapter 7.)

* The need to make predictions and cope with mistakes.
In making choices and pursuing its motives an organism needs to predict the consequences of possible actions and the behaviour of other parts of the environment. Our environment does not seem to be completely predictable on the basis of human intelligence, though there are some useful regularities. Even when events are in principle predictable, the available information, reasoning abilities, or time, may not be adequate for making reliable predictions. Moreover, information relevant to making or correcting predictions may come at unpredictable times. This implies a need for constant or frequent monitoring of the environment, and an ability to notice and deal with the unexpected, in a rule-governed fashion, instead of always blindly executing previously available instructions: another example of the need for decompilation. In particular, it may be necessary to be able to interrupt or modify both mental processes and physical actions to cope with sudden dangers or opportunities, as in the example of fright. Sometimes, it will suffice to re-route, alter speed or manner, or change a not yet attempted subgoal, whereas at other times an action will need to be aborted, for the sake of a very important motive which was previously dormant.

* The need for speed.
Speed of computation and action may often be important. Events in the environment may happen quickly, relative to the speed at which an organism can perceive things, make inferences, take decisions or move. The organism may sometimes have to move quickly relative to the speed at which it can acquire and process information, make plans or take informed decisions. This is why it may be necessary sometimes in the light of new information to interrupt an action in order to deal with some unrelated motive.

The need for speed can influence the trade-off between space and time in the choice of representations, since the brain has a limited finite speed of operation. In particular it may be important to have many rapidly accessible pre-computed strategies, rules of thumb, consequences of known facts, etc., instead of always working things out from general principles. This is because search spaces in deriving useful results from general principles tend to be subject to combinatorial explosions. (Imagine always using Peano's axioms as a basis for answering arithmetical questions: we find it essential to store lots of 'partial results' for rapid access.) This necessitates extra space for information storage, and good indexing schemes. A corollary is that information is stored in a highly redundant form. The difficulty of acquiring complete and correct

information means that mistaken generalisations may be used to derive a great many different consequences, and even if the mistake is discovered later it may not be possible to track down and modify all the factual beliefs, procedures, goals, etc. derived from it. Although it is desirable in principle to store reasons for decisions and beliefs (Doyle 1980), it may not be possible to be completely systematic about this, or to use the information about changed reasons to work out exactly how all the consequences should be modified.

So it is to be expected that the system will be complex, messy and not necessarily consistent in its information. When they have slow brains intelligent systems will tend to have, and develop, very large and messy information structures. (This may seem to support "scruffy" AI system building (Abelson 1981). However it is no support for a lack of clear theory about such scruffy systems).

In some cases action may be so urgently required that there is no time for normal processes of inference, deliberation and decision, and it may be necessary to 'short circuit' them. Mechanisms for doing this could include reflex arcs at the level of hardware and very high-priority interrupts at the level of software. Since speed will be important on some occasions and not others, it will be necessary for the system to be able to distinguish different degrees of urgency and importance associated with motives.

* The need for improvement.

Almost any skill or ability is capable of indefinite improvement, in several different dimensions, such as speed, precision, the variety of conditions in which it can be applied, reduced undesirable-side effects, degree of effort required, the ability to combine it with other skills, etc. Mechanisms for extending and improving skills, as well as detecting and removing mistaken beliefs are desirable. Thus failures and successes (especially unexpected successes - a sub-case of prediction failure) need to be analysed to see what lessons can be learned from them for future efforts. The extent of decompilation, i.e. explicit representation of information and strategies, will help to determine how much informed learning is possible. A compiled (implicit) strategy is hard to modify intelligently.

* The need for motive generators.

Because the environment may permit particular needs to be satisfied in a variety of ways (e.g. because food may take many different forms), a general goal is often not specific enough to produce action. New information needs to be acquired concerning what is available in the environment. This information should be capable of interacting with the general motive to produce a more specific version: e.g. a desire for food may be transformed into a desire for a particular edible object. We shall see presently that there is a need for a variety of types of motive generators, including some which react to the social environment by absorbing tastes, principles, etc.

* The need to cope with changing motives.

'Body monitors' are needed if the organism or machine has a complex body. Many different sorts of disturbances and malfunctions may be possible, including excessive external forces, a temperature which is too high or too low, shortage of fuel or lubrication, intrusion of foreign matter, breakage, etc. The system will need many monitors capable of detecting such events and either causing automatic corrective action (if urgency is great and no intelligence is required) or

causing a new motive to be added to the system's store, possibly associated with a high priority so that actions generated by other motives may be interrupted.

Old motives can also generate new ones as subgoals for achieving some task. Moreover, new motives can arise from internal processing, e.g. inferring a new danger, on the basis of previous knowledge. Mechanisms for changing the current set of motives are required. This also may make it necessary for the system to be able to interrupt and re-direct processing or actions, since new motives may be incompatible with and more important than old ones.

* The need to plan over a long time scale and interleave actions.

Some opportunities for achieving goals, or avoiding dangers, exist only during a relatively short time interval. Information, food, materials, etc. may become available at times when they are not needed, and be unavailable when they are needed. The ability to store objects and information is therefore important, and the organism needs to be motivated to use the ability. The camel's hump, the squirrel's hoard, and the human brain all address this problem. More generally, since causal influences may involve long time delays, it may be possible to perform an action which contributes to satisfying a motive only if it is done a long time in advance. For instance, to shelter in a house during the winter you must start building it long before. You can eat crops only if you've done the ploughing and sowing long before. It is thus necessary to be able to *interleave* the achievement of different intentions. This implies an ability to store information about partially executed plans and to notice conditions for continuation. This, as we shall see, is one of many reasons why processes concerned with one motive may interfere with other processes.

* The need for motive comparators.

As we have seen, there are many reasons why a complex organism, especially a social organism, can be expected to have a large and complex collection of changing and not necessarily mutually consistent motives. Choosing between alternative goals, strategies, styles of performance, etc., will not be a simple matter. The notion of an *optimal* choice will not necessarily even be well-defined (Sloman 1965). Achieving a long term balance between different needs of the individual or the larger community can be a major problem.

Scarcity of resources or opportunities, or the existence of various hostile agents, or the existence of goals concerning the well-being of others, may make an individual's needs and goals mutually inconsistent. There may even be intrinsic (logical) inconsistency between different motives with different origins in the same individual: for instance the same person may have sexual motives and motives absorbed from a monastic cult. Motive-generators and reasoning procedures derived from erroneous beliefs may remain after the factual errors have been corrected, as hinted above in discussing the need for speed. This can also lead to inconsistent motives. Mechanisms and strategies for detecting and dealing with such inconsistencies will be needed. It is therefore necessary to be able to relate the consequences of different motives and make choices between them, including choices which allow long term motives to override short term desires. In short, 'motive comparators' are needed. These may be compiled into automatic priority mechanisms, or decompiled into explicit rules for choosing in the light of available information. The more explicit the comparators, the easier it is to alter them in the light of their consequences: an important form of learning.

* The need for different modes of execution.
Margins of error in the performance of actions are very variable, as are the risks associated with error, so that different kinds of care and attention may be required in different circumstances. Walking on a broad parapet requires less precision than walking on a narrow wall. Errors on a high wall are more serious than errors on a low wall. The ability to compute margins of error and potential dangers, and trade the need for speed against the need for care and precision may be important in some environments. Compare Sussman (1975) on 'Careful mode'. This is a case where instead of *interrupting,* the action produced by other motives, an important motive, e.g. staying alive, or avoiding pain, *modifies* the mental and physical processes involved in performing the action. For instance, the action may be slowed down, or done with more precisely controlled movements, or the results monitored in more detail. Other kinds of modifications of plans or actions may enable two motives to be pursued simultaneously, for instance talking loudly in order to convey information to one person and attract the attention of another.

* The need to cope with other agents.
Sometimes the actions or attitudes of other agents matter (for instance because they are potentially co-operative or hostile), so it may be important to perform actions in a manner which gives information to others, or conceals information. In humans, many motives (some very powerful) are concerned with influencing what others believe or feel, since this can, in turn, influence how they act. The ability to have and act on such social motives requires the ability to represent the mental states of others. This plays an important role in many human emotions.

In conditions where resources are scarce and self-directed motives powerful, it seems that only if individuals are provided directly with the desire to please and help others (as opposed to treating this as a mere means to self-satisfaction) can various sorts of mutually destructive conflicts be avoided. (E.g the prisoner's dilemma.)

This is an example of a more general point. In co-operative communities, it may be important that individuals develop motives which do not necessarily maximise their own advantage, but which enable the community as a whole to function well. This could apply to teams of robots as much as to animals.

* The need for motives not to be rigidly pre-programmed.
The fact that conditions can change enormously (partly as a result of social evolution, partly through 'natural' causes), implies that if genetically programmed information cannot be rapidly revised then the genetic information about the environment and strategies for coping in it should be restricted to what is of relatively permanent value. Other information should be learnt, either directly by individuals, or indirectly from other knowledgeable individuals, and represented in such a way as to be changeable: another example of decompilation. Hence the need for very high level motive generators, and mechanisms for 'absorbing' tastes, rules, principles, values, etc. from the current culture, or developing them in the light of experience, instead of or in addition to genetically programmed 'drives'. To allow adequate response to changing circumstances, any tendency to conform may need to be supplemented by motives which generate experiments with alternatives to the norms of the community. This can, of course, lead to conflicting motives within and between individuals.

NOTE
The need to learn motives (including tastes, aesthetic and ethical principles, standards of behaviour, etc.) could apply to machines which are intended to be capable of functioning, in collaboration with human beings, in a wide variety of cultures whose details cannot all be known in advance, or which might change during the life of the robot.

* The need for motives relating to the helpless young.
Individuals may perish, and new ones be produced. (The need for this may arise in part from the increasing difficulty of reorganising knowledge systems as new discoveries are made and old assumptions revised. Beyond a certain point it may be best to start with a fresh new mind and give it the latest information only). Depending on the normal life-span, it may be important for the transmission of information to offspring to be much quicker than the original discovery process. Lamarckian mechanisms (if they exist) and more overt communication both raise similar problems of how information may be converted to a form suitable for transmission to another agent, and how it can be decoded. (The representation within a working system may be unsuited for communication to another system: for instance, printing out computer data-structures in an intelligible form is often a non-trivial problem.) Mechanisms and strategies for such communication must be supplemented with motives which ensure that they are used. The task of communicating skills and knowledge to the young may be a considerable burden to the older individuals, and conflict with many of their motives. It is therefore necessary for additional motives to be somehow programmed into them (whether genetically, or through learning, or both) to ensure that they do the job: yet another source of non-selfish motives.

Our list of constraints defines a set of questions to be asked about organisms: do they cope with the constraints and if so how? For the robot designer, they specify design tasks which he may or may not undertake. Not all the constraints will be met by all animals. Empirical research is required to establish which animals have which abilities. For extinct species, many of the questions may be forever unanswerable.

It is essential to our main argument that in human beings
(a) information is always incomplete and partly erroneous
(b) reasoning processes take a significant amount of time, and are error prone
(c)there are many different sources of potentially incompatible motives.

Being able to cope with (c) in the light of (a) and (b) i.e. in a complex and partly unpredictable and uncontrollable world that we cannot fully understand or keep up with, is central to our notion of human intelligence.

There is no reason to assume that all motives derive from one over-arching motive or drive, or even a small number, or that all decisions are based on the goal of optimising something measurable called ‘utility’. Even where we can theoretically perceive the unifying purpose, or function, of a cluster of motives, there need not be any explicit representation of the purpose in any individual, nor any explicit derivation of the other motives from it. For instance, the representation and derivation may be implicit in the process of natural selection, or in the mind of a programmer.

Moreover if an individual's collection of motive generators and motive comparators is the result of a long and complex process of trial and error learning (by the species, the culture, or the individual) then his system of motivators need not satisfy minimal canons of rationality (e.g. preferences may not be transitive). It is particularly true of motivators absorbed from a culture, or genetically transmitted, that they need not have survived because they are of particular use for the individual.

Paradoxicallly, then, intelligence as we understand it does not rule out ignorance and irrationality. Rather it involves being able, at least to some extent, to recover from the consequences of such limitations. We attempt to explain how, in what follows.

The computational architecture of a mind

Can we derive mechanisms from the design constraints listed above? Chapters 6 to 10 of Sloman(1978) sketched some of the computational architecture of the human mind in the light of common sense knowledge and task analysis. Some of the details were left very vague. What follows is an attempt to fill in a little of the missing fine structure. The constraints don't completely determine the design, but do suggest some general features.

An intelligent system needs a 'central' administrative process concerned with deciding: forming intentions, making or selecting plans and resolving major conflicts in the light of motives Up to a point, this might be implemented as a hierarchy of different parallel processes, with different levels of authority, and different areas of expertise (like a business organisation, or army). However, since not all decisions can be taken independently, and sub-processes will sometimes generate incompatible goals, it will be necessary for some conflicts to be resolved at a high level in the light of major goals, policies, etc. (Simpler alternatives, such as fixed priorities for different subsystems might be found in less intelligent organisms.)

It was conjectured in Sloman (1978) that the distinction between what we are and what we are not conscious of is primarily a distinction between what is and is not, at any given time, accessible to and representable by this 'central administrator'. (Dennett (1979, pp30,ff) claims that this notion cannot account for the 'privileged access' one has to one's own consciousness. In Sloman (1974), pp.302,ff an alternative explanation of privileged access is based on Frege's sense-reference distinction. We therefore see no reason to follow Dennett in relating consciousness to the 'speech centre'.)

In what follows we show that the central adminstrative process needs to be thought of as a number of asynchronous sub-processes, if the constraints of the previous section are to be met. Instead of a fixed program, with well-defined temporal orderings, maximum flexibility in dealing with surprises is achieved by decomposing the tasks into parallel sub-processes, some of which can interrupt others. Thus instead of each program having built into it conditions under which it must pause and allow a decision to be made as to whether it is time to attend to some other business, the decision can be taken by some other, dedicated process. This helps to ensure that opportunities are not lost: like the examples of decompilation discussed above, this allows important decisions about sequencing to be taken at 'run time' as opposed to 'program construction' time. It also allows different sub-processes to run at different speeds according to their own needs. This would be hard to achieve with an essentially sequential control mechanism.

Among the necessary sub-processes we distinguish processes producing motives, a process of selecting motives for consideration, a process of choosing a subset of the considered motives as a basis for action (forming intentions), a process of deciding which intentions should be acted on when, a process of selecting or constructing plans for achieving intentions, a process of executing plans, and various monitoring process. (See the Appendix.) How far these processes are distinguished in humans and other animals, and how far they are able to run in parallel is an empirical question: we can suggest, but not demonstrate answers.

The essence of parallelism is a certain unpredictability and the prevention of missed opportunities in one process because of the continued running of another. This can be simulated as closely as required on a single serial processor, provided it is fast enough, and has a suitable operating system. Simon (1967) argues that in humans there is a single resource-limited serial process, which is capable of simulating parallelism. Sloman (1978, chapters 6 and 8) and Sloman (1981) argue that there are both empirical facts and design considerations which suggest that different sub-programs should be capable of being run in parallel under the control of other programs. For instance, if a monitor can interrupt another process, then the latter need not *guarantee* termination, or even fairness in releasing resources to other processes. This is a commonplace of operating system design. Even Simon agrees that at some level of description different things go on in parallel.

If there are many different sorts of internal processes at any one time, concerned with controlling different movements, interpreting sensory information, making decisions in the light of new information, etc., any change of plans may involve very complex changes, as sub-processes are killed, indexes updated, records of what has and has not been done revised, and so on. Internal monitoring may produce awareness of some of these processes. We try to show below that awareness of such disturbances, is part of what characterises emotional experiences. (This is a generalisation of the older claim that emotional experience is based on awareness of physiological changes.)

Motives play a central role in our discussion. Simon (1967) uses the term ‘motivation’ to designate ‘that which controls attention at any given time’. This definition is too vague. For instance, the execution of a learnt skill may involve attention being directed to a variety of objects and activities, without any *explicit* purpose. It seems therefore that Simon’s definition is too general, since any step in a program would be a motive. It is easy enough to give examples of the narrower sense of ‘motive’, e.g. desires, dislikes, preferences, ideals, ambitions, etc. But a general characterization is harder.

Without departing too far from Simon’s definition, we define ‘motive’ to refer to representations, conscious or unconscious, which generate internal or overt behaviour, in conjunction with other representations playing the role of beliefs and strategies. To complete the definition, we must embed it in a theory showing how motives beliefs and strategies interact. Calling motives representations implies that they are symbolic structures with descriptive content. (I.e. they have a semantics.) There need not be any awareness or experience of having the motive. Motives produce behaviour in several different ways. So we need to survey the types of processes in order to understand the nature of motives. Part of the difficulty of giving simple and clear analyses of these and other concepts of mind is that we are dealing with closely interrelated functional concepts, and the definitions must therefore be mutually recursive. Complete definitions are possible only with complete theories of the nature of mind (at least at a certain level of description). So, we can only hope to approximate ever more closely to adequate definitions via a spiral of ever more complex and complete theories of mind.

In accordance with the principle of promoting flexibility by decompilation, we shall find a need for additional entities which we call *motive generators* and *motive comparators.* We use the term 'motivator' as a general label for all of them.

In the next section, we sketch some of the main processes underlying decisions and actions, to provide a framework for more detailed analysis of types of motivators in a later section. We have already seen that there are several types of choices: choosing which motives (e.g. desires) should become intentions, choosing which intentions to act on at any time, choosing how to achieve them, and so on. We shall observe repeatedly that for maximum flexibility in a partly unpredictable world, there is a need for processes generated by one motive to be able to interrupt or modify processes produced by others. Such flexibility is a hallmark of intelligence. If we can show that such mechanisms are also capable of generating emotions, this will establish that the ability to have emotions follows from meeting the design criteria listed earlier, which did not mention emotions.

Processing motives

* Adding and removing motives.
Most fundamental are the processes which create new motives. For instance body monitors and motive generators (defined below) may add motives, and the process of planning may produce new subgoals from old goals. Simon (1967) mentions a number of ways in which sub-processes may be terminated: achievement, satisficing (partial but adequate achievement), impatience and discouragement. However, there are additional ways in which motives might disappear. A physical desire may disappear because something has happened to the body, such as a chemical change or illness. A motive which is never acted on because others are always chosen instead, might be caused by a 'forgetting' mechanism eventually to disappear. A desire which always leads to suffering when acted on might therefore be eliminated by an inductive learning mechanism (see below). Higher order motives might be capable of banishing motives from the store as well as introducing them. Interrupting a process of pursuing a goal may cause subgoals to disappear. A subgoal can be removed when the main goal has been satisfied 'accidentally'. The plan to board a train may be abandoned because the person you were going to see has just stepped off it. This can also happen when the subgoal is a postponed intention, which has not yet generated any action. Notice that removing redundant motives presupposs computational requirements: recording of reason-motive links. (cf. Doyle 1980 and the 'purpose-process' index of Sloman 1978.)

* Selecting motives to act on (forming intentions).
Since there may be a large and not necessarily consistent collection of motives, it is not always possible for all of them to be acted on. Some sort of selection process is therefore needed. Motives selected for acting on will be described as *operative.* Rejected motives are *inoperative.* (Some may then be removed from the system.) Motives which are neither accepted nor rejected, but left for further consideration later will be described as *suspended.* The ordinary word 'intention' covers only operative motives. 'Desire' includes motives which may be operative or inoperative. Even insistent desires can be over-ridden by other desires, fears, etc. or higher-order motives such as aesthetic or moral principles, or the social motives referred to previously. (Desires do not go away just because they are over-ridden.) Suspended motives are often referred to when we say we have not yet 'made up' our minds whether to do something (not

when or *how* to do it!) This may occur because there is not yet enough information to decide which to select, or because measures of desirability, or merit, or rules for comparing motives, fail to yield a clear cut decision, or because there has not yet been enough time to consider the motive.

The process selecting operative motives may make use of a variety of different criteria, including for example, urgency, insistence, merit (desirability), difficulty. These are defined below.

Urgency is easiest to define, being concerned essentially with how much time is left for the motive to be acted on. Of two motives which are equal in other respects, it would normally be sensible to choose to act on the urgent one first. Deciding which is most urgent may involve arbitrarily complex cognitive processes.

Insistence is a measure of the ability of a motive to attract the attention of the selection process. It is a measure of how hard it attempts to get itself selected, i.e. presses for attention. (This measure could be represented explicitly by some symbol associated with the motive, or implicitly in the state of a processor.) Common motives with high insistence include those derived from bodily needs. Other insistent motives can be part of emotional states, for instance an insistent desire to make someone suffer for what he has done. In computationl terms, insistence can be thought of as an interrupt priority level. The process interrupted is the administrative sub-process concerned with choosing intentions (operative motives). This makes sense only if the system separates the process of motive selection from the process of motive formation: perhaps less intelligent animals do not. A motive may be very insistent without being selected for action, as when a seven year old child recently said: 'I have an itch which is very eager to be scratched': only later did he abandon his game and remove his shoe in order to scratch! You may have a desire to cause suffering to an individual, yet choose not to do anything about it, even though you have the opportunity and frequently experience the inclination. Hopes, ambitions, idle wishes, etc. can be present for a long time without receiving any attention: their insistence is low.

Insistence of a motive is distinct from *merit* which may be based on arbitrarily complex assessments of consequences, preconditions, etc. in the light of motive comparators. The point is that the degree of insistence of a motive is its power to be *considered* for selection -- something like an interrupt priority level -- whereas its merit, or desirability, is its power to be *selected* for action, as against other incompatible motives of equal urgency or difficulty. It is not immediately obvious why there is a need for insistence as well as merit to play a role. One reason may be that the process which selects motives is resource-limited, so that if it spends its time constantly cycling through all the inoperative motives looking for ones worth promoting to intentions, then it may take too long to complete each cycle, and something important and urgent may be missed, especially as some decisions whether to adopt an intention can involve arbitrarily complex computations. In that case, greater flexibility, and more frequent re-assessment of needs may be achieved if the mechanisms which create new motives are allowed to give them a quickly computed prima-facie assessment of importance, represented as an interrupt priority level, i.e. what we have called insistence. Something with high insistence may then cause the motive selector to consider it sooner than it would otherwise do.

This seems to illustrate a general principle concerning the division of labour in a hierarchically organised intelligent system: don't leave all the important decisions to the highest level process, including the decision whether to re-direct its attention.

Insistence and merit are independent. Something which continually interrupts and gains attention may be continually rejected by the selection criteria. Sitting on a sharp object may produce a very insistent desire to arise, that is a desire that is very hard to ignore, though this may be regarded as having far less merit than other motives, such as not giving away one's whereabouts. Conversely, something not very insistent or urgent may frequently win the battle for selection. One may regard certain duties as very meritorious though the motive to perform them does not constantly intrude in one's decision-making. Here is another example of the tension between compilation and decompilation. In a well designed system the insistence measures will correlate well with the result of a fully informed merit computation. One kind of learning could be the improvement of procedures for assigning insistence levels, so as to reduce the frequency with which the motive selector is interrupted to give attention to a new motive with lower merit than those currently under consideration. A frequently computed merit result for a class of motives might be compiled into an automatically assigned insistence measure for that class. However, this may prove harmful if circumstances change. The long-term usefulness of such inductions is one of several dimensions of cognitive friendliness in the environment.

In a 'super-intelligent', fully informed system there need never be a conflict between insistence and merit: the priority assigned by the low-level processes to new motives may conform exactly with their relative merit. Ensuring this would require all processes to have full access to all the knowledge, reasoning powers, etc. of the main administrator. Full knowledge in advance is presupposed, for otherwise some lower level processes concerned with assigning intensity levels might need to be capable of chaotically interrupting more global processes, e.g. by redirecting eyes, or sending the whole organism off to get new information! Such anarchy might be tolerable for some organisms in some environments. In a machine, or animal, with limited knowledge and computing resources, it might be better to have the functional division in which insistence measures are computed very quickly, with limited resources, to give a rough and ready estimate of urgency and merit. Some kinds of mental disorder may turn out to be due either to malfunctions of the mechanisms for assigning insistence levels, or to the learning of inappropriate rules for doing so.

The merit attributed to a motive considered for selection may but need not be linked to effects of achieving or not achieving it. A 'derived' measure of merit of a motive could be based on the frequency with which pursuing it facilitates satisfaction of *other* motives. For instance, acquiring knowledge, or health, may help one achieve many other goals. However, some things may simply be valued more highly than others: motive comparators, whether inherited or learnt, may directly assign a higher priority to one motive than another, for instance, when reading a book is 'simply' preferred to having a meal. The degree of insistence of a motive may also enter into assessment of merit. This is particularly desirable where insistence is controlled by mechanisms which measure objective and relatively urgent needs, like mechanisms which produce a desire to drink.

Difficulty is another measure which can play a role in selecting motives. It is related to time required for some task, to the physical effort required, to the amount of prior knowledge or skill involved, to the complexity of planning required, to the number and variety of obstacles to be overcome, and to the presence of conflicting motives with high insistence. (Monica Croucher has begun to analyse the notion in an unpublished paper.) Sometimes difficulty can outweigh merit. Something with very great merit may remain inoperative, or postponed, on account of the effort, or time, required to achieve it.

The consequences, urgency and difficulty of a motive do not conspire to define a unique meaningful measure of priority. So decision rules, rather than some kind of summation, are needed. Such rules, for instance, may specify how to weigh up some short term pleasures against long term annoyances, or vice versa. There is no reason to believe that a single general principle can be used by humans or other animals as the basis for such rules. Instead, *motive comparators* may have to be developed on a trial and error basis, possibly under the influence of social processes. They constitute heuristics for choosing between different sorts of motives when conflicts arise. These may involve arbitrarily complex computations, for instance gathering information about the effects of possible actions. If built up through a lengthy process of trial and error they may not be totally consistent, for instance when preferences are not transitive.

* Deciding which intention (operative motive) to act on when.

In a system with a variety of independent motives, the selection process may form more intentions (operative motives) than can be acted on simultaneously. So a time-ordering will have to be chosen. To some extent this may be based on intrinsic characteristics of the motives. Some may require special conditions which will not exist till later. Some may require several steps separated by long time intervals. Some motives may be explicitly conditional, e.g. 'If it rains take steps to keep dry', or 'Whenever danger approaches, run the other way'. Thus, there are several different reasons why not all operative motives can be acted on as soon as they are chosen. They will need to be temporally ordered, or possibly interleaved. This requires computational resources for keeping track of what has and has not been done, and when things are to be done.

In principle there might be systems which lacked the ability to do this long term planning. They would be inherently less flexible. Informal observation suggests that different people have this ability to different degrees, and that very young children may lack it. This raises interesting questions about how such an ability might be learnt.

Intentions (operative motives) which are not currently being acted on (although they have been selected for action) will be described as *dormant,* the others *active.* Some dormant motives are concerned with a specific action, and simply wait for a suitable opportunity to perform that action, whereupon they are fulfilled, for instance waiting for the ice to freeze and then skating. Others are more *general* motives, concerned with a class of conditions and the actions to be performed whenever they arise. When the relevant circumstances turn up, such a motive generates more specific intentions taking account of the circumstances. (This is not the same as generating a sub-goal. Intending to get food, and leaving it till later to decide which kind, is not the same as adopting the subgoal of going to a supermarket to get the food. A specification is different from a means.) Many traits of character or personality are like this. For instance, most of the time you are not in the process of trying to do anything to help people in distress, yet if you observe someone have an accident, the new information can interact with a dormant but operative general motive to generate a new specific active and operative intention to help the sufferer. An ambitious person will have many operative dormant motives waiting for opportunities to generate more specific intentions. Thus a dormant general motive is a type of *motive generator,* which can interact with new information to produce a specific motive.

Contrast an *inoperative* motive such as is to be found in some marginal cases of hypocrisy. The motive to help others may actually be present and frequently considered as a basis for deciding what to do, yet always overridden by other motives. Behavioural tests would be unable to confirm its presence.

A dormant motive may have been explicitly postponed, to a definite time, or to a definite stage in the execution of some other intentions ('I'll cut the grass as soon as I've finished sharpening the mower'). This presupposes some kind of timing process or explicit manipulation of plans which will ensure that the motive is activated at the right time. Alternatively, it may be postponed until the right conditions occur, at some unpredictable time. This requires a 'demon' mechanism: a monitor which will detect the occurrence of the conditions, and cause the motive to be activated, possibly disturbing other actions.

The priority ordering, and selection of active motives among intentions may make use of processes very similar to those which select intentions from among motives in general. Urgency, difficulty, merit, opportunity may all be taken into account. Relative merit may be determined by motive comparators, such as a preference for eating fish over fowl, or a moral principle that gives preference to helping those least able to help themselves. Jon Cunningham has pointed out (privately) that the same sorts of *mechanisms* might be deployed to select motives to attend to, and to select motives to act on, but using different criteria, and effectively operating at different levels within the global organisation. This is consistent with a view that evolution of complexity is partly a matter of new combinations of old mechanisms.

* Drawing attention to suspended or inoperative motives.

There are several different reasons why suspended or inoperative motives may be re-considered and become operative. Conflicts with other motives may have been removed, either because the other motive disappears, or because circumstances change so that they are no longer incompatible. (Learning about contraceptive techniques for instance may remove an inconsistency between motives.) Alternatively higher-order motive-comparators may change their criteria. (Mixing with a different social group can lead to a new view of the merit of chastity.) An increase in insistence may cause an inoperative (i.e. previously rejected) motive to be reconsidered, for instance a decision to do nothing about a pain may have to be changed if the pain gets much worse, even if a high-level motive comparator rules that it would be best to ignore the pain.

The reconsideration of suspended and inoperative motives could be achieved by a variety of mechanisms, including frequent re-examination of the whole set of inactive motives, or, more plausibly, by some mechanism whereby the motives themselves compete for the attention of the motive selecting process. Each suspended motive might be assigned a processor which constantly runs an appropriate monitor program, which decides when it is appropriate to interrupt the selector process. Some inoperative motives may not do this until circumstances change and wake them up. Others may constantly request attention, like an itch.

* Changing the insistence of inoperative motives.

Physical needs (e.g. shortage of fluid, damage to tissues, etc.) may be monitored within the body by mechanisms able to insert appropriate motives into the store. To allow urgent needs to be dealt with quickly, they should be able to alter the insistence of motives, that is their ability to interrupt the intention selection process.

Other more 'cognitive' processes may also be capable of altering the insistence of motives, for instance the desire to annoy someone may be inoperative (e.g. because over-ridden by moral scruples), yet may grow more insistent as a result of his actions. This could be the result of a general inductive process which creates, or makes more insistent, a motive to harm another agent who continually harms oneself. This sort of motive generator could play an important role in

social regulation. We shall see later how it can play a role in anger.

Sometimes, owing to new information, the merit as well as insistence of a motive will change, in which case, on being reconsidered, the motive may be selected for action. In the case of physical needs, the insistence of the motive tends to be correlated with the objective merit, except perhaps in young children, addicts, compulsive eaters, and pathological cases.

* Waking up dormant motives.
A decision to help people in need, or to run away when predators approach, will not be effective unless perceptual processes are available for recognising people in need, or predators, and activating the relevant motive. This requires communication between perceptual processes and the store of motives and motive generators. The data-driven mechanisms required for dormant motives need not be the same as those used by *active* motives, which may require goal-driven visual searching or checking, involving complex high-level processes which cannot be made available for routine monitoring.

If all monitoring is done by a single relatively slow time-shared process, then there will be a limit to the number of dormant motives which can be properly served. If each such motive can be assigned a different processor, which runs in parallel with all the others, then the number might be larger though it is likely that the processors, and therefore the *type* of monitoring will have to be limited in complexity. In people it seems that there are some upper limits, but that the kind and variety of monitors can be influenced by training. In some animals it may be that there is a fixed stock of monitors, genetically determined.

* Making or choosing plans for execution.
Operative active motives will require actions to be performed. In some cases a suitable procedure will already exist, and can be found in the store of resources. In other cases, it may be necessary to construct a plan. This may be an arbitrarily complex process, including for example, attending courses or reading books to gain relevant information. (This sort of process has already received considerable attention in AI.) Sometimes alternative plans will be available, or constructable, and relevant motives may have to be invoked for choosing between them (e.g. a desire to maximise speed, minimise dangers, minimise expense, and satisfy certain aesthetic criteria).

Finding alternative plans then assessing them is less efficient than letting preferences control the process of construction, so that instead of being rejected after construction, bad alternatives are not even constructed. This can be achieved by permitting global motives to be 'consulted' during the construction of a plan. If the set of potentially relevant goals and preferences is small, then this can be done by the planning process frequently running through the set of potentially relevant motives, looking for clashes, opportunities to 'kill two birds with one stone', etc. (This was implied in Sloman 1978, chapter 6). Where the set of motives is large, such frequent polling would take too much time and would interfere with the planning process, in which case some parallel monitoring process, comparing current partial plans with relevant motives, and interrupting the planning if necessary, would be more efficient. This requires internal perceptual processes which interact with motive generators just as external perception does.

The pressure to save time may cause some motives and motive comparators to be compiled into planning procedures. Instead of always referring back to motives the planning process, like any other well learnt skill may use previously compiled procedures for choosing. Some global preferences may be transformed into local strategies for suggesting extensions to partial plans.

Others may be transformed into local monitors set up during the planning process to object to certain partial plans at an early stage. Thus motives may be 'compiled' into proposers and filters used in planning. So what starts as an explicit motive (a requirement for intelligent systems) may be turned into an implicit one embedded in an planning mechanism. This can save time and improve performance. Thus global intelligence may be enhanced by compiling some intelligent processs into unintelligent, automatic ones. The cumulative effect, if circumstances change during a long learning process, may be to produce a messy inconsistent system which works quickly, but does not always work well.

* Executing plans.
This may generate overt actions or internal processes, like adding new subgoals to the store, making plans to fill in gaps found during execution, removing subgoals when they have been achieved, etc. Considerable amounts of book-keeping may be involved if plans are at all complex. In a system made up of several processors, it may be possible for a plan or sub-plan to be executed independently by a lower-level process, allowing the central administrator to carry on with more difficult tasks, such as creating new plans. This devolution will be possible for plans which have been constructed previously and thoroughly tested (Sloman 1981). If the lower-level system is also capable of taking decisions, forming sub-plans, etc., it too will have to have a store of motives, or access to some or all of the main store.

* Detecting success and failures.
Processes of perception play a major role in execution. Dead-reckoning will not usually work in a complex and unpredictable world, so constant monitoring will be required. Sometimes things will go unexpectedly well, e.g. because new opportunities or resources turn up, or obstacles go away, and sub-goals and sub-plans can then be abandoned. Sometimes the reverse happens, and new plans have to be formed, actions repeated, etc. Where the action is mostly mental (e.g. the goal is to form a plan, the action is building one), the perceptual processes will be purely internal: checking on the relationships between the plan, the goal, and available information. Such internal perceptions may use the same mechanisms and strategies as the higher levels of external perception.

Successes and disasters may occur at surprising times, in an incompletely understood world. If they are to be detected, it is therefore necessary to permit the relevant perceptual processes to be driven more by the data than by the current plan of action. Hence, in an intelligent and flexible system perception and comparison of new percepts with current motives and motive generators should occur in parallel with action, and be capable of interrupting action. Simpler systems may only be able to alternate between performing a sub-action and performing a pre-determined test, like the 'TOTE' units of Miller et al (196?).

* Retrospective analysis of success or failure.
Failures sometimes provide information from which the organism can learn how to do better next time. So can unexpected successes. An intelligent system will need to analyse the relationship between the satisfied or violated motive and what happened, so as to be able to find out how to do better. Even actions which go exactly as expected may sometimes provide new information during execution. Analysis of the sequence of events may reveal unnecessary redundancy, missed opportunities, previously unnoticed risks, and possible generalisations or specialisations, which should in future be taken into account, etc. So retrospective analysis is always potentially useful. For this reason it will be useful to have a built in tendency to look back at and try to learn from

what happened during the execution of an action. This might be an explicit motive or compiled into attention-directing mechanisms, which react to failure. This sort of learning requires analysis of the relation between procedures followed, the various goals and subgoals involved, and what actually happened (E.g. see Sussman 1975). Some systems may be able to do this retrospective analysis in the course of performing extended actions. They would then be able to improve their performance 'on the fly'. Other systems may have to rely on simply storing records of what happened, and doing the analysis at times when nothing else is going on. Perhaps this has something to do with sleep. Retrospective analysis may involve hypothetical exploration of alternatives to what actually happened: a possible interpretation of dreams.

Where parallelism is restricted, the tendency to reflect on past events may have to compete with other motives for attention. This competition should be controlled by some kind of priority system, which assigns a priority to the analysis. Such a mechanism might account for some emotional states in which a tendency to dwell on a previous success or disaster interferes with present needs.

Some desires may cause considerable disturbance if left unsatisfied. For instance, lack of food can lead to general weakness and an inability to pursue other motives. We shall see that more general mechanisms can also cause unfulfilled motives to produce disturbances. If information about effects of fulfilment or non-fulfilment is available to the system (e.g. as a result of inductive learning), it can contribute to the assessment of the merit of such desires. The disturbance may be physical (as in the case of some addictions, and some genuine physical needs), or mental, for instance causing other motives to be violated by preventing adequate concentration on achieving them. This is a source of some kinds of emotions as we shall see. An individual need not be a good judge of how much he will suffer from carrying out a task, or from non-fulfilment, or, indeed, how much pleasure or satisfaction will result from fulfilment of the motive.

* Inductive generalisation.

Besides producing new factual beliefs, inductive learning can produce new motives, motive generators, motive comparators, skills and strategies, and modify old ones. Sometimes such learning is a result of *understanding* what has happened previously (Sussman 1975). However, a relatively *blind* inductive learning process would be useful when causal connections exist which are not intelligible to the learner. If doing something is frequently and consistently followed by satisfaction of other existing motives, then it might be useful to develop a new motive to do that thing whenever possible, or increase the insistence or strength of an existing motive (or motive-generator). Where insistence implicitly represents such past successes, it may be useful for the merit computation to take account of insistence.

Similarly, if a type of action or situation is frequently followed by the frustration of other motives (e.g. unpleasant things happen) then it might be useful to develop a new motive generator to avoid that sort of action or situation, or lower the insistence or strength of an existing one. Blind induction occurs when the connection between the action and its results are not understood. (What understanding is, requires further analysis!) The existence of this sort of mechanism would explain some results of behaviourist research. But it presupposes powerful cognitive mechanisms.

An interesting trade-off arises here. If the learning process directly modifies motives, or their priorities, without recording the evidence, then this speeds up subsequent decision making, but possibly makes it less flexible. If, however, instead of *compiling* a changed motive, the system stores the relevant information and uses it later to weigh up relative priorities, then this will slow down decision making, but allows greater freedom, for instance permitting other considerations to outweigh the records of previous results in special circumstances. (The general strategy of recording *reasons,* is defended at length in Doyle (1980). In a system with limited storage or speed, it can be overdone.) Probably the best compromise is to store reasons for a while, and then to recompile the motives or motive generators after several occasions of use if there have been no conflicting considerations. At that point the reasons can be discarded, or perhaps merely become inaccessible. This exemplifies an important form of cognitive development, in which results of previous intelligent inference and problem-solving are compiled into new 'unanalysed' unintelligent operations.

* Persistence in failure

A spider or thermostat which fails to achieve its goal may go on trying the same moves over and over again -- and with luck will eventually succeed. Most puzzle-solving or game-playing programs will give up after the first failure. Clearly the motivation to persevere also has merit in more intelligent systems since repeated actions will not always produce the same result in a partly unpredictable world. However, it is also important to be able to judge when to give up or try something different. Computer operating systems often use a simple count of the number of tries. A more intelligent system would be able to vary the number of tries in the light of experience, and would sometimes not use counting as a criterion at all. Perseverance and flexibility then should both be reflected in the strategies of an intelligent system.

* Interruptions and parallelism.

If one administrative sub-process can occur in parallel with other processes, then new results may cause the first process to to interrupt the other processes. This allows greater flexibility than a system which allows new information to be considered only when the single central adminstrative process temporarily abandons other activities. This illustrates a recurring theme about the greater flexibility and modularity of parallel processes which permit interrupts, as compared with processes in which everything happens in accordance with synchronised plans. Most work in Computer Science on languages for distributed systems seems to favour the latter restriction on account of the difficulty of proving properties of systems which permit interrupts. This suggests that if the human mind has the sort of architecture proposed here, then human mental processes may be theoretically very difficult to predict, even when full information is available about the programs involved.

* Impulsive vs deliberate interruptions

It is possible for people to be trained so that opportunities or dangers produce automatic actions, for instance a boxer noticing his opponent has slightly lowered his guard, or a member of a committee noticing an opportunity to make a point. Going through the full process of deliberation and planning may take so long that the opportunity to gain or avoid something is lost. It is as if the processes which create such motives are able *directly* to invoke a previously learned strategy for acting on them. Alternatively, perceptual processs may be able to use some kind of associative mechanism to trigger actions directly, totally bypassing the representation, comparison and selection of motives. (Notions of moral or legal responsibility can founder on

such possibilities.)

The mechanisms making possible this kind of 'short-circuiting' could play a role in some kinds of emotions as well as impulsive behaviour. It is perhaps curious that in the simplest goal-directed systems this relatively inflexible procedure may be the only one available, whereas in a more intelligent system, able to choose between motives and strategies, the more primitive mechanism may have to be reinstalled to enable emergencies to be dealt with properly, on account of limited processing speed. Once again, decompiled mechanisms need to be speeded up with additional compiled mechanisms.

* Suppressing interruptions.

A system with many parallel sub-processes capable of interrupting other processes, might be totally chaotic. One way in which the chaos could be reduced would be to prevent the interrupt mechanisms being totally automatic. That is, they might be decompiled into rule-governed mechanisms which, instead of merely interrupting take note of information about whether an activitiy is interruptable or not. The simplest way to achieve this is to have interrupt thresholds: for instance, a motive whose insistence is below some threshold may not be permitted to claim attention. The threshold might be varied according to the importance and urgency of current high level goals.

Sometimes a motive, such as hunger, or the need to deal with physical discomfort, appears not to get awakened because full attention is given to something else: this could either be due to limited parallelism, or to modification of interrupt thresholds in some processes caused by activity in others. For instance, when a plan relating to a motive of high merit requiring full attention is being executed, it would often be advantageous to disable certain classes of interrupts, to ensure that only motives relevant to the current activity are capable of being considered by the selector. This is analogous to what often occurs in a computer operating system, or during garbage collection in some languages. This may be similar to the reason why people who are injured in battle, or while playing in a football match sometimes don't notice the injury till long afterwards. In people, it seems that this mechanism can be used at times by motives which objectively do not have very high merit, as when someone totally engrossed in daydreaming or reading a novel fails to smell the smoke in the next room, or hear a child's question about why the baby has turned a funny colour.

The inability of relatively insistent desires, pains, etc. to get any attention at all during the performance of some actions, might seem to imply that not all the sorts of processes sketched above run in parallel on independent processors. Alternatively, it may be the case that what we are conscious of is restricted to information accessed by some one process, all of whose resources are needed for certain tasks, while other processes may continue in parallel, on other processors. The third possibility, just mentioned, is that some specially important activities can alter thresholds so that interruptions are prevented.

* Speeding up deciding.

Inability to decide between alternative motives or strategies for achieving intentions can lead to disasters. Opportunities can disappear before the decision is reached. It may therefore be necessary to be able to detect when deciding is taking too long and generate a motive to speed up the decision-making process or perhaps terminate it.

* Random processes.
It may be useful to have some kind of randomising mechanism which is occasionally allowed to select between alternatives. For instance at times when it is not possible to make a reasoned decision it may be urgent to take *some* decision between equally plausible alternatives. If there are enemies about who are intelligent, they may be able to learn the decision making strategy if it is not random, and use it to their advantage. However, it is not necessary that the mechanism be as random as one based, for example, on quantum phenomena would be. It is sufficient (as Phil Agre pointed out to me) that inferring the strategy be beyond the reasoning powers of the 'enemy'. Random strategies in conditions of uncertainty may maximise the opportunities for learning.

In organisms which possess a randomising mechanism there is always the possibility that it may be invoked when further reasoning about pros and cons would be more desirable. Teachers often get the impression with some students who are having great difficulties in learning some technical concepts or skills, that they invoke a randomiser as a result of a strong desire to provide *some* answer quickly, so that they fail to use their real abilities.

In the light of the design constraints listed earlier we have begun to explore some of processes that may be useful in intelligent systems and some of the mechanisms capable of supporting such processes. We have seen that there are different ways in which one motive can interrupt or modify processes generated by another. New or previously inoperative or suspended motives may need to interrupt a process deciding which motives should become intentions. A new motive becoming operative, or detection of conditions relevant to operative but dormant motives can interrupt the processes of planning or execution. These can also be interrupted by detection of new circumstances relevant to the current action.

Some of the interrupts may be capable of being handled by a lower-level specialist subsystem executing a relatively simple plan, for instance a process of moving a hand out until it touches something. Some reflexes are concerned with such local control. Other interrupts require global re-organisation. Some interrupts may involve matters of such urgency and merit that the new motive bypasses normal selection and planning processes and directly causes a (previously learned) strategy to be invoked. This is most common when sudden dangers are detected (for instance, detecting that you are about to lose your balance, or that something is crashing down on you).

Besides *interrupting* (Simon 1967), a motive may *modify* an ongoing process, for instance causing it to be done more quickly or slowly, or in a 'careful mode', or using a different set of tools or sub-goals. It is suggested in (Sloman 1981) that the synthesis of new skills from old may require some procedures to be executed under the control of others, which can vary their speed and other parameters. We shall see that some emotional states, for instance anxiety, may involve these sorts of interactions between different motives. Such mechanisms, which are required for flexibility in coping with a complex and unpredictable environment in the light of a complex set of motives, may also be capable of generating dysfunctional processes. Not all emotional states are useful.

When an individual contains a complex collection of motives and motive generators, with possibly delicately balanced conflicts between mutually inconsistent motives, the system is likely to be very sensitive to experience, with relatively small external changes producing new motives, decisions and actions, which in turn can lead to further internal changes. (Such systems might

easily get into unstable states.) Thus individuals who start off very similar can develop in very different ways.

This in turn will make it necessary for members of a community to develop great sensitivity to variations in individuals, since crude general rules will not be reliable for predicting behaviour. (The difficulty of making such predictions might generate evolutionary pressure towards mechanisms which make individuals show their attitudes and feelings whether they want to or not. Human facial expressions, posture, tone of voice, crying, laughing, etc. seem to perform this role.)

This sketch (summarised in the Appendix) of some of the possible processes involving motives in an intelligent system provides a framework within which we can understand the need for different sorts of motives, concerned with different roles in these processes. It suggests that human decision-making can have a complexity not reflected in our ordinary language. We often speak of deciding or choosing, as if there were one type of process, whereas it should now be clear that there are very different sorts of selection to be distinguished, especially selection of motives to become operative, selection of operative motives (intentions) to be active at a particular time, and selection of means of achieving or fulfilling active motives. Further, we see a need for many parallel processes, monitoring internal and external events for occurrences which are relevant to the possibility of initiating new processes, which may mean interrupting old ones. We do not claim to have established that any or all of the processes and mechanisms described are to be found in human minds. That requires elaborate empirical research, and no doubt it will be found that not all human minds are identical. Similarly, not all this complexity will be found in all animals, or all intelligent machines. So far, we have tried merely to provide a theoretical framework to guide empirical research as well as the design of intelligent machines. Further, by more detailed examination of the implications of different ways of implementing processes of deliberation and decision, we contribute to the long-term goal of exploring the space of possible intelligent systems.

Having surveyed some of the processes generated by motives, we now look a little more closely at some of the different sorts of motives, elaborating on distinctions already hinted at. After that we can turn back to emotions.

Types of motivators

'Motive' was defined above to refer to a variety of types of representations capable of generating or controlling processes of the sorts described in the previous section, namely processes of choosing, planning, acting, etc. The simplest type of motive is a goal, a specification of something to be achieved, prevented, preserved, or terminated. The wider class of motivators includes motive comparators and motive generators. Motivators subsume what we ordinarily call desires, dislikes wishes, tastes, preferences, ambitions, intentions, principles, ideals, attitudes, and some personality traits, like generosity or selfishness. A more detailed analysis of these concepts would show many hidden complexities, such as that talk of someone's *likes* may either refer to what he desires, or to what he experiences when the desires are fulfilled, or to what he thinks the experience would be.

Motivators need not be stored separately from beliefs, procedures, etc. In fact, for reasons of efficiency it may be necessary to store some of them in association with concepts or schemas used for recognising objects or situations. This is reflected in ordinary language when motivating adjectives like 'dangerous', 'pleasant', 'trustworthy', 'inefficient', which are concerned with

reasons for acting or choosing, play the same grammatical role as purely descriptive adjectives, such as ‘red’, ‘round’, ‘unstable’, ‘flexible’, ‘sticky’. Thus a store of beliefs may contain pointers to information relevant to forming intentions or choosing between alternatives. This motivational information may be attached to representations of *individuals* as well as types, e.g. when an object acquires sentimental value, or a person is a friend or enemy or loved one. This integration of factual and motivational information may be crucial to the implementation of close links between the perceptual processes and the activation of motive generators, dormant motives, and the like.

What makes a representation a motivator is not where it is stored or what its structure is, but its role in processing. In some systems the role may be determined by location, in others by form: that is an implementation detail. It is even possible that some motivators may be ‘compiled’ into other representations -- beliefs, decision making procedures, habitual strategies, the choice of descriptive language, etc., so that there is no longer an explicit representation of the goal or preference. Many of the motives we attribute to people (ambition, pride, generosity) may in fact be implicit in mechanisms of decision and deliberation, or distributed implicitly among other lower-level motivators. It is particularly hard for such ‘motives’ to be recognized, explicitly compared with alternatives, or removed from the system.

There is no implication that a person, animal, or robot which has motives is conscious of all the motives, or even of all the processes by which motives generate decisions and actions. For instance, the detection of implicit ‘compiled’ motivators would require very sophisticated self-analysis. (Sometimes other observers can see the pattern unwittingly produced in one’s behaviour.) Self consciousness requires both internal perceptual processes which build descriptions of what is going on, and also accessibility to such descriptions by the central administrative process. (For a brief account of computational reasons why some mental processes are not accessible to consciousness see chapter 10 of Sloman 1978.) One feature of implicit motives may be that failure is not detected: without an explicit representation of a desired state of affairs the inadequacy of the actual state of affairs may go undetected.

In the simplest case, an explicit motive will simply be a representation of something to be achieved, prevented, preserved or terminated. We often use the word ‘goal’ to refer to such simple motivators. However, as already indicated, more complex sorts are possible. We can now summarise some of the kinds of variety to be found in motives.

* Varying logical structures.
Motives may have any of the logical complexity of factual statements. That is they may include negation, disjunction, conditionals, universal quantification, existential quantification, and so on. For instance, one may want either not to have company at all or to have the company of someone intelligent (i.e. someone unspecified). When the logical structure of a motive is complex, it may be difficult to decide whether it has been achieved or violated. Aesthetic preferences and religious beliefs can apparently generate motives such that it is hard to distinguish achievement from non-achievement (e.g. "Do God’s will"). Some motives may be embodied in non-propositional representations, for instance a painter’s specification of what he is trying to paint, the musician’s internal representation of the tune to be produced.

* Basic and derived motives.
Some motives are present only because they are derived from others, and would be abandoned if the others were abandoned or fulfilled. For instance, some are simply sub-goals of other motives.

Achieving them need not be thought of as desirable: there might even be a preference for finding a way of avoiding having to achieve a particular subgoal. In other words, not all goals are valued intrinsically. Whether subgoals are eliminated when they are no longer required may depend on the thoroughness with which purpose-process indexes are maintained (Sloman 1978 ch 6).

Sometimes a motive which was originally subservient acquires an independent status, and may subsequently function as something intrinsically valued. How this sort of thing happens is not clear, though there is plenty of evidence that it does happen in people. It may sometimes be a result of the sort of inductive process described previously, where the effects of acting on some motive or motive generator cause its own priority to be modified. The important point at present is that whether something plays the role of a basic or derived motive simply depends on how it happens to be represented and related to other motives. How it originated may no longer matter. Thus something may become intrinsically desired which does not serve any useful function like preserving life, improving health, etc.

* First-order and second-order motivators.

This is the distinction already mentioned between specific goals (first order) and the higher level motive generators and motive comparators. A goal produced by a motive generator, may have the status of a basic independent desire. For instance, a "benevolence" motive-generator may produce a motive to help an individual in distress. Once produced the new motive can function as a strong desire, whose insistence or merit is not directly dependent on contributing to achieving some other goal (utility, pleasure, avoidance of pain, etc.). Part of educating children to be civilized is instilling such motive-generators in them. Many environments instead instill racial or religious bigotry and intolerance: a different set of motive-generators. There may also be higher-order motivators, for instance motive-generator-generators and motive-comparator-generators. Thus a ‘conformity’ motivator could cause the individual to absorb not only specific goals from the community but also more general rules for generating and comparing motives. Some of the second-order motivators in humans and other animals may be the result of evolutionary pressurs concerned with the needs of a group, or a gene-pool, rather than the needs of an individual. Examples would be motive generators in adults concerned with their reactions to helpless infants, and sexual motive-generators.

* Dimensions of ‘strength’

We have seen that motives can differ in insistence, in urgency, in merit, and in the effects of satisfying or violating them. Merit is not a one-dimensional scale, but may be a complex and perhaps even inconsistent partial ordering implicitly defined by a large and not necessarily consistent collection of motive comparators, built up through a long and erratic process of learning (phylogenetic and ontogenetic). Thus theories of decision making based on scalar representations of strength of motives may be inaccurate and unable to account for some of the richness of human and animal behaviour.

* Instincts and learnt motives

Some motives and motive generators may be genetically programmed, such as (perhaps) a motive generator which produces desires to find things out: a ‘curiosity instinct’. Others may be produced by motive generators, including body monitors. Motive-generators and motive-comparators may themselves be either innate or learnt. Our discussion shows that the genetically programmed motive-generators may be very varied sorts of mechanisms, some producing motives very directly (body monitors), others only as a result of complex cognitive processes,

such as retrospective analysis and inductive learning, or the perception of distressed infants or potential mates.

* Alarm bells versus directives

Some motives merely indicate a need for something to be done, leaving it to other processes to determine what the problem is and how it is to be solved. Many sorts of pain and discomfort have this character. The representation constituting the motive may include information about the *location* of the trouble, or about its *seriousness,* even if the precise nature or cure is left unspecified. Sometimes the problem is actual damage (e.g. punctured skin). Sometimes it is merely potential damage (e.g. a high temperature, or a heavily loaded muscle). Mental states may also produce relatively non-specific motives: a vague unease, or discomfort, indicating a need for something to be changed, but leaving open what. Such motives require the expressive power of something equivalent to the existential quantifier. In the simplest cases, mere movement of the relevant part of the body will remove the problem. In other cases far more elaborate processes may be needed, including diagnosing of the source and producing a specific motive to deal with it, e.g. taking some medicine.

* Pleasure and displeasure

Body monitors (hardware), or mental motive-generators (software) may also generate motives to *preserve* some state or process. These may involve physical sensations, physical or mental activities (e.g. sex, eating or drinking whilst not enough has been consumed, the process of taking in some kinds of information or developing some skill). There may be operative but dormant motive-generators watching out for such situtations, then planting motives to keep the state or activity going. 'Pleasure' and 'enjoyment' are words we often use to characterise this sort of thing. As with pains and discomfort, some motive generators concerned with pleasure may be genetically programmed and relatively permanent, whilst others develop over time. An activity which at one time is not enjoyed may become a favourite pastime, perhaps as a result of the kind of inductive learning already described. It might be useful to have very general motive generators which produce desires to achieve tasks found to be difficult but manageable, or to complete tasks once they are started. It may also be useful to have mechanisms for reducing the insistence or merit of certain motives when attempts to fulfil them require disproportionate amounts of effort, or when evidence has been found that the goal is unachievable. (This would be connected with some emotions, e.g. discouragement.)

Pleasure and displeasure may be associated with success and failure of actions. When it is learnt through perceptual processes that an attempt to achieve a goal has failed, the unfulfilled motive will tend to generate remedial planning and action at the same time as the a different process produces a tendency to analyse the failure and learn from it. The original specific motive and the more general learning mechanism may therefore combine to produce a tendency to dwell on the failure, which, in turn can produce an emotional state. The existence of a desire for the situation to be different from what it is, is at least part of what is involved in finding the failure unpleasant. Similarly, when a goal has been achieved, especially if it was not easy, there may also be lessons to be learnt, and a tendency to reflect on the achievement would be useful. The desire to go on dwelling on the success in a certain fashion is part of what is involved in gaining pleasure from success. These remarks on pleasure and displeasure hardly scratch the surface of a cluster of very complex concepts.

In the case of social organisms, both success and failure are likely also to be associated with actual or potential effects on other members of the community, and awareness of this will influence the processes which result, as in shame and pride.

Decision-making cannot be completely rational

As remarked previously, motive selection processes may involve measures, such as insistence, merit, urgency, difficulty, likelihood of success, pleasure or pain produced; and may also use descriptive rules (motive comparators). The various measures and comparison rules may define only a partial ordering of motives, leaving some cases of conflict unresolvable, since neither motive clearly has more merit than the other. If there is a strong motive to reach a decision between the two incommensurable motives, the violation of that motive may produce considerable disturbance: one sort of emotion. Similarly if one motive is selected as having greater merit, the other may retain its high insistence (perhaps because it is connected with some sort of warning mechanisms which it would be too dangerous ever to be switched off) and continue to demand attention, another kind of emotional state.

The processes by which decisions are taken are sometimes thought of as analogous to mechanical processes of the kinds studied in physics. For instance we talk about 'weighing things up', about being 'pulled in different directions' or 'balanced on a knife-edge'. However, it should be clear that mechanical metaphors are inadequate to represent the complex interactions between different motives, at least in a human being. For instance, two physical forces pulling an object in different directions produce a resultant which may pull the object in an intermediate direction with a force which is stronger than either of them if the angle between the two directions is acute. However, motives are not additive in any such simple mathematical fashion: when confronted with a choice between two palatable dishes you don't grab at the space between them. Plato found it more fruitful to postulate internal processes analogous to the governing of a city. His metaphor begins to match the computational richness required, but, since it assumes intelligent voters, debaters, etc. it risks circularity. The circularity will be avoided when we have specific designs for computational systems capable of doing the tasks.

One of the problems, however is the difficulty of specifying what the task is, when so many complex motives interact without any well defined concept of what constitutes the best decision. Some cases may be dealt with by very general strategies, e.g. choosing at random if all the conflicting goals are of roughly equal merit, or choosing the urgent one if one of them must be achieved soon or not at all whilst opportunities for the other will recur. Even where the degree of merit is represented by a measure, this may be very coarsely quantised, leaving a need for additional heuristics for selecting between goals with the same measure. There may even be a collection of different, incommensurable measures, for instance if one scale is used for assessing intellectual pleasures, and other gustatory pleasures. Some of the problems of choosing might be handled by a strategy of cycling the priorities, so that different sorts of motives take turns at being given highest priority. However, as already remarked, the notion of an optimal decision, and therefore of an optimal decision-making system is probably not well-defined. (For more on the complexity of comparing alternatives see Sloman (1965).)

Since there are so many different types of motives, with so many different types of origins (including different environmental constraints and the needs of social groups), there is no reason to assume that all motives are explicitly or implicitly derived from or able to serve any one general principle or motive, like maximising chances of long-term survival, or maximising

happiness, or minimising pain. Mathematical models of decision making based on such assumptions, e.g. a consistent partial ordering of possible outcomes of actions may be inadequate to represent most human and animal decision-making processes. Even if all decisions had to be related to the ultimate goal of maximising some measure, then the computational task of doing this accurately would probably be quite intractable in the time normally available before opportunities for action are lost. Strategies for coping with lack of time for decision making would then have to be taken into account. These could be dependent on previous experience and vary enormously from individual to individual. The search for general 'laws' of decision making would then be fruitless. Perhaps a few culture-based regularities might emerge. All these remarks may be applicable not only to organisms, but also to a 'general purpose' robot designed to integrate well with a variety of human communities.

Could a machine have its own motives?

One of the issues surrounding the debate whether machines could ever think, or whether it is appropriate to use mentalistic language non-metaphorically to describe machines, is whether machines could ever be said to have their own motives. Many people cannot accept that this is possible. This may be partly because machines already known to us either are not goal-directed at all (clocks, type-writers, cars, etc.) or else, like current game-playing computer programs, have a very simply structured hierarchical set of goals. Asimov's 'laws of robotics' illustrate such a hierarchy. There may be a single goal which generates subgoals in the light of a variety of constraints. The highest-level goal is there simply because the programmer put it there. If however, we were ever to design machines with a system of motivators (including motive generators), with the kind of complexity described in previous sections, then whatever the set of motives built-in at the time of manufacture, the machine would develop and change over time in such a way that some of the original motives would be completely over-written, and new motives, generators and compartors could be produced which flatly contradict many of those originally provided. In that case it would be totally misleading to say that the machine was pursuing the goals of its designer. We could, of course say that although there were goals, preferences, principles, generating decisions and actions they had no owner. It would, however, seem much more reasonable to ascribe them to the machine, in just the same way as we do to people.

However, there is no logical argument to compel anyone to talk this way: for ultimately the decision whether to do so is a *moral* decision, concerned with how we ought to treat the machine. People who for religious or other reasons refuse to allow that a machine could have a right to be given consideration and kindness, will persist in requiring that only biological origins (Boden 1977) or some other criterion which rules out machines, can justify saying that an individual's purposes are its own.

Methodological comment.

The characteristics and components of intelligent systems listed above should not be taken as a *definition* of intelligence, for there is no single set of necessary or sufficient conditions for describing an organism or machine as intelligent, conscious, having a mind, etc. Rather there is a whole spectrum of cases, from systems with very simple feedback loops to those containing all the complexity sketched below. Arguing about where to draw the line between cases of real intelligence or mentality and the rest is quite pointless, like arguing over whether it is still 'really' chess if one player starts without a queen. What is not pointless is the analysis of

different cases and their properties. The impossibility of drawing a sharp dividing line has ethical significance, but that will not be discussed here.

Our approach needs to become more interdisciplinary. In order to identify the sort of computational system required to model the human mind it is useful to start by trying to make explicit what we already know about what sorts of human processes can occur. Chapter 4 of Sloman(1978) outlined a strategy employed by philosophers, called "conceptual analysis", for articulating knowledge implicit in ordinary language and thought. Some of the contents of this essay result from application of this strategy. Conceptual analysis is not to be confused with introspection, for the process involves considerable trial and error, like any scientific activity. (It has a lot in common with methods used by Linguists: and has similar limitations.)

It has to be supplemented, in the manner we have illustrated, by investigation of some of the constraints within which organisms or robots may have to function, and which determine what would make them well or ill adapted. This is familiar as ‘task analysis’ in AI, though it is not often attempted on the gobal scale begun here. In addition, an empirical survey of patterns of processing in different sorts of animals may help us understand some of the sub-mechanisms in more advanced species. (See Lorenz (1977) for the beginning of such a survey.) Psychological research on skills, errors of performance, emotions, effects of brain damage, etc. may contribute tests for alternative postulated mechanisms.

It is difficult to find out a great deal about the inner workings of a computational system simply by observing or experimenting on it. Attempting to design something similar on the basis of detailed task analysis may be more informative, at least until good theories are available with testable consequences. What has been outlined above certainly has consequences, though many of them are not novel. Moreover, as in much of science (Sloman 1978, chapter 2) the predictions concern not so much what will happen as what can happen, like a grammar which does not enable a single utterance to be predicted yet predicts that a very large number of utterances are possible. More detailed analysis of ways in which the mechanisms can go wrong could lead to predictions and explanations of phenomena involving mental illness, learning difficulties, etc. It might also provide some predictions concerning ordinary emotional states.

Emotions: an example

We have now laid the ground-work for an analysis of some of the processes involved in emotions. Many of the processes described above could occur in intelligent beings with physiological processes totally unlike ours, even beings in whom some emotions were not correlated with any bodily changes (except the brain-processes required to support the computational processes).

To illustrate some general themes, we look at a common emotion, anger, in some detail. There are different types of anger, but we initially consider only what might be called the ‘standard’ case. Typically if some individual X is angry, there will be another individual Y with whom X is angry, and something about which X is angry. It may be that Y did something X wanted him not to, or that Y failed to do something X wanted done. It need not be the case that either of these *actually* occurred. All that is necessary is that X *believe* something like this was done by Y. Either way, the object of the anger is believed to be responsible for something which violated one of X’s motives. This in itself is not sufficient for anger. All of this could be true, yet X might merely wish the occurrence had not happened. He might *forgive* Y, and not be *angry* about it.

For X to be angry, he must in addition want to do something to hurt or harm Y, i.e. something which violates one of Y's motives In other words, part of the anger is a new motive in X directed against Y. This new motive might be produced by a general-purpose motive generator (innate or learnt) which reacts to situations where another agent frustrates one of X's motives. (If sufficiently general, this might produce anger directed at oneself on occasion.) The new motive may be very insistent, yet not regarded by the motive comparators as having relatively high merit. Instead of being selected for action: it may remain inoperative, for instance because X is afraid of the consequences of hurting Y, or because some other motive overrides the desire. Alternatively, it may be selected as an intention, and so become operative.

The mere existence of the desire to harm Y, together with the other circumstances, is not sufficient for X's state to be one of anger. He may have the desire, but put it out of mind, and calmly get on with something else, and in that case he would not be angry. Anger normally involves an *insistent,* desire to do something to Y, that is, the desire should frequently 'request attention' from X's decision-making processes. So the desire will frequently come back into X's thoughts, making it hard for him to concentrate on other activities. This aspect of anger, and other emotions, makes essential use of the 'interrupt' mechanisms already shown to be desirable in an intelligent system able to cope with the constraints mentioned earlier. It is also possible for the new motive to *modify* other processes, instead of simply interrupting them. For instance, the processes of selecting other motives, strategies, or executing plans, may be modified in such a way as to alter their effects on Y. "Cutting one's nose to spite one's face" describes a class of such cases. Thus the manner in which X performs various actions, the precise means selected, and so on may be chosen so as to harm Y, or minimally to avoid pleasing Y.

When the strength of the emotion has died down, and can be put out of X's mind, there may still be long term effects, which probably result from some of the learning mechanisms which we saw earlier are desirable in intelligent systems. For instance, for some time afterwards, anything which reminds X of Y may also remind him of the occasion Y did whatever caused the anger. Remembering the previous occurrence may in turn regenerate the disturbed state, in which the desire for revenge becomes insistent and disturbs other processes. This sort of reminding could be of use in a world where things which have harmed you once may do so again.

Rational anger and be distinguished from irrational anger. The conditions listed so far do not suffice for rational anger. For example, the desire to harm Y must not be one which X would have had in any case: it must arise out of the belief that Y has violated one of X's motives. That is, X's *reason* for wanting to harm Y must be his belief that Y has done something X objects to. So, if the desire to harm Y is not redirected to Z on learning that it was Z, not Y, who was responsible, then the anger is not rational. There are other types of irrationality, for instance such as desiring to do something to Y which is disproportionate to what Y has done, or which will cause more harm to X than to Y.

Different forms of anger and related emotions are distinguishable according to some of X's second order beliefs about his motives and actions. Righteous anger includes a belief that moral standards, or some kind of external authority was violated, and not just one of X's motives. This sort of anger requires sophisticated concepts which many animals, and very young children, lack.

We have already mentioned one kind of disturbance which anger can produce, namely constantly intruding into X's decision making. It is also possible, in human beings, and probably other animals, for the anger to produce physical disturbances, such as sweating, shaking, feelings of tension, tendencies to perform violent actions, such as stamping, thumping objects. Where

there is *self-monitoring,* the perception of such states of the body will play a role in determining the total experience. These bodily disturbances probably also make use of mechanisms which are required for other purposes, for instance rapid alterations of physical actions. However, it is not *conceptually* necessary that anger involve any such physiological effects. If X satisfied all the other conditions he could rightly describe himself as being angry, possibly even very angry, despite not having the physical symptoms. The social implications of his being angry might be the same with or without the physiological changes. Even without physiological effects the feeling of anger could be strong, insofar as it constantly intruded into his thoughts and decisions, and insofar as he strongly desired to make Y suffer, and suffer a great deal. The reasons why other people are interested in your anger have little if anything to do with your physiological state or your awareness of it. The have a lot to do with your beliefs and intentions and future plans and whether you can control some of your motives. Often the physical effects are counter-productive insofar as they give information to Y. Good teachers, and, presumably, spies and diplomats, learn to control some of the physical effects. (They may also learn to control some of the mental processes, e.g. learning not to *become* angry in certain conditions.)

Although we have stressed that anger can interrupt and disturb other activities, it need not actually do so. For instance, X may be engaged in pursuing his new motive of hurting Y. Here the activity is produced by the anger: the anger is not therefore interrupting some other activity. The characteristic of anger, and other emotions is a *disposition* or *tendency* to interrupt, disturb or modify other activities. The disturbance may be entirely mental, of course: we are not talking only about external behavioural dispositions (compare Ryle 1949).

The strength of anger, like other emotions, can vary along different dimensions. It can vary according to how much X minds what he thinks Y has done, which, in turn, will depend on the 'merit' (in the sense defined previously) of the violated motive. The strength can vary according to how much harm X wants to do to Y. (Many an angry parent wants to smack a child without being at all willing to see the child tortured horribly.) The strength can vary also with how much merit the new wish to harm Y has: the desire may be very hard to override, or it may be relatively easy, though not necessarily easy to forget (i.e. it may lack merit -- in our technical sense -- though it is insistent). Finally, the strength of the anger can vary according to how much general disturbance it is capable of producing in X, whether physical or mental.

By considering different combinations of conditions, we see how anger is related to other emotions. When there is no desire to cause harm to Y, the emotion is more like *exasperation* than anger. If there is no attribution of responsibility, then the emotion may merely be some form of *annoyance,* and if the motive that is violated by some event or state of affairs has very high merit, and cannot readily be satisfied by some alternative strategy, then the emotion is *dismay.* (These are some of the types of variation to be taken into account in a complete taxonomy of emotions.)

The analysis so far has described some minimal conditions for the state of being angry. In addition to *being* angry, one may *feel* angry, though this is not inevitable. For instance, a person who is emotionally involved with another may be unaware of the fact, even though others recognise the state, as novelists and playwrights frequently remind us. Feeling the emotion requires processes of self-monitoring. This seems to be what is involved in what we call 'having the experience' of anger. The state may in some sense be perceived, and yet not described or classified (like the visual experience of looking at some totally unfamiliar complex machine). This would be a different experience from the sophisticated self-awareness which labels mental states as emotions. Different animals, or robots, and even different people may be capable of

different degrees of precision in the internal self-monitoring. 'Experience' is a very complex concept, which is not easy to analyse, and further work is required to clarify the issues. Some of the states and processes involved in anger might be conscious, others unconscious. In particular, even if information about internal processes is accessible to self-monitoring processes, they need not have the descriptive resources to characterise what is happening accurately, so there may be a felt emotion which is not recognised. We have noted that it is possible to be angry whilst totally unaware of the fact: *being* angry does not entail *experiencing* anger. However, if it is very strong (in the senses described above) then the disturbances it produces will generally be noticed. If a person is conscious of only certain aspects of his state, he may then misclassify it.

Towards a more general account of emotions

This is not yet a complete analysis of anger, but it does reveal a number of features which can be shown to be common to a wide range of emotions, including fear, embarrassment, guilt, joyful anticipation, sorrow, disgust. We can now sketch a first list of typical features or components of emotional states, different combinations of which would generate a taxonomy of emotions. (Some of the features may also be present in non-emotional states.) Some aspects of the analysis were developed independently by Roseman (1979), though his scheme is not rich enough in generative power: it admits only a relatively small number of types of emotions because it does not allow for the variety of interactions of multiple motives, nor the possibility of recursive escalation enriching emotions.

* Being in an emotional state involves having at least one fairly strong initiating motive Mi (with high insistence and 'merit'): e.g. a desire for something to be the case or not be the case, past present or future. Real or imagined or expected satisfaction or violation of this motive or set of motives produces a disturbance or modification of other goal-directed processes, or a *disposition* to disturb or modify such processes. (In the latter case there is not actually a disturbance: for instance the disposition need not gain control. NB: the disturbance can be fruitful, as when fear of failure produces more careful and attentive action.) Sometimes the disturbance will produce a new motive Mn, such as X's desire to harm Y resulting from the initiating belief Bi (see below) that Y has violated Mi.

* Having an emotion (e.g. being angry) is not the same as experiencing it. The individual concerned may or may not be aware of the disturbance or potential disturbance, depending of the sophistication of internal monitoring abilities. *Feeling* or *experiencing* the emotion seems to require such internal perception. If there is an internal representation of the state, this self-awareness can activate further dormant motives or motive-generators, for instance producing a second order motive Ms, such as a desire to control the emotion, or not to show it, or not to have the motive Mn. Violation of this new motive Ms can produce a further emotional state superimposed on the first: recursive escalation.

* There must be some initiating belief-state Bi concerning the initiating motive Mi: e.g. a belief that it has or has not been satisfied or violated, or will or will not be, or explicit uncertainty about the matter. This generates several sorts of cases, depending on whether the motive is *for* something, or *against* it; whether the desire is thought to be satisfied or violated; or whether there is uncertainty about which is the case. A further distinction can be made as to whether the belief Bi concerns the past or the future.

a. The belief that something undesirable has happened can generate despair, the belief that it will happen can generate desperate efforts to prevent even what is thought to be inevitable.

b. A belief that something desired has happened can produce an emotional state involving considerable pleasure or satisfaction, while the belief that it will happen can produce a desire to make doubly certain that it will by taking additional precautionary steps. Depending on other beliefs and motives, this may be a state of anxiety, or of joyful anticipation.

Believing that something has or has not happened -- or will or will not happen, is involved in emotions such as anger, sorrow, despair, shame. Believing that something could happen or fail to happen is involved in anxiety, hope, anguished expectation, and so on.

In some cases the belief Bi will be concerned with two or more initiating motives - for instance a belief that they cannot be simultaneously fulfilled. This can produce a desire Mn to overcome the contradiction, or to find the best way to choose what to give up. Various sorts of anguish may result.

* For an emotion to exist, the combination of motive(s) Mi and belief Bi (or in some cases uncertainty) must be capable of producing a considerable *disturbance in activities concerned with other motives and beliefs.* This could arise from processes which adjust the insistence of motives, and which generate new motives, such as Mn and Ms, in the sorts of ways described previously. So one consequence of the belief Bi about violation or satisfaction of the motive(s) Mi, may be an inability to completely put the topic out of one's mind, and perform other mental tasks normally: the new motives Ms and Mn may have high insistence, so that the tendency to dwell on what has happened or will happen, continually intrude into other thoughts and decisions, and influence one's perceptions, unless some other, important motive, generates urgent action in such a way as to raise interrupt thresholds, in the manner described above. (Unconscious emotions may do all this in a fashion which is evident to other observers, though not the sufferer.) In some cases people seem able fruitfully to redirect emotional states like anger into work on something quite unrelated. How this happens still needs to be explained: talk of 'channelling energy', or 'redirecting' merely describes the phenomenon without explaining it. Perhaps it could be explained by inductive processes previously referred to, which produce changes in the operation of some motive generators and processes controlling insistence of motives.

Examples of new motives Mn with high insistence (i.e. high interrupt priorities) include a desire to right what has gone wrong, a desire to avenge a deed, a desire to inform other people, a desire to gain the sympathy of others, a desire to repeat whatever it was that pleased one, a desire to take additional steps to make sure of success, a desire to gain additional evidence about what has happened or will happen, and so on.

The disturbance need not involve a new motive such as Mn or Ms: there may simply be a constant dwelling on what has happened. We have indicated why a tendency to dwell on successes and failures may be an important part of the process by which an intelligent system learns. In an emotional state regrets or happy recollections may constantly intrude in one's

thoughts. Where new motives are generated, we may call the emotion *active,* otherwise *passive.* (Jon Cunningham has remarked that this implies that organisms lacking certain learning abilities will therefore lack the ability to have certain emotions.)

In some emotional states there are not just one but several mutually inconsistent initiating motives Mi, for instance when a person does you and your loved one's several kinds of harm, and also performs some action you greatly admire.

* As already mentioned, the new motives need not be selected for action: they may remain inoperative, though with a constant tendency to disturb or modify other processes. By contrast, some emotional states such as fright may involve the direct production of actions, by-passing the process of forming an intention and selecting or constructing a plan for executing it. That is, the motive whose violation is perceived or anticipated may be able directly to activate some already existing strategy, interrupting whatever other actions were in progress. This would include cases of 'impulsive' action. The mechanisms making this possible would also make it possible to take very rapid remedial action in times of great danger, or when sudden opportunities are recognized. (In less intelligent organisms, this direct production of actions by motives may be the only way they generate action.)

* Some emotional states arise out of actions performed by the individual. For instance, as pointed out in the discussion of general constraints on intelligent systems, some actions require great precision, and the risks of error may be very serious. In this case, during the performance of the action there may be considerable fear about possible errors, and secondary motives generated to take extra care, suppress interruptions, monitor actions in detail, etc. These secondary motives may, in extreme cases, generate so much disturbance of the risky action (e.g. frequent thoughts about whether enough care has been taken, attempts to devise alternative strategies, etc.) that they can reduce the speed or precision with which the action is performed, and lead to the very disaster they are concerned with avoiding. An over-anxious driver would be an example.

* One dimension of variation between emotional states concerns the speed with which things happen. We have mentioned several times the importance of interruptions in an intelligent system coping with a complex and partly unpredictable environment. In some cases, the disturbance produced by an emotion is partly a matter of interrupting a large number of ongoing processes, for instance processes controlling different parts of the body, and rapidly redirecting them to cope with some new danger or opportunity. If the relevant parts of the body contain sensory detectors, then the many local changes produced by the interruptions will be monitored, and the system's perception of its own state will be changed. This sort of emotional state does essentially involve physiological processes. However, this kind of experience is not a necessary part of all emotions for all intelligent systems.

* The emotional state may have been deliberately sought, as when people go to see a sad play, or seek out thrills. In this case there need not be actual interruption or disturbance of other activities, though there will often be physiological disturbances which are experienced. Yet the potential to interrupt other processes remains. Attempting to turn ones attention completely on to trying to prove a theorem, or plan one's summer holiday would be difficult. If it proves easy, then the emotion may have been simulated, and the screams or tears fraudulent. Sometimes special training can make it easier: as if new mechanisms make it possible either to change the interrupt

thresholds, or alter the interrupt priorities of disturbing processes. Sometimes in people an emotional state alters cognitive abilities in a fashion which has nothing to do with the cognitive content of the emotion. It may be that physiological processes, especially chemical processes in the brain, play a mediating role (perhaps increased adrenalin improves some cognitive skills). But if so this is a contingent fact about how humans function, not a necessary feature of emotional states.

* The existence of some ability to detect and represent the internal state does not necessarily imply that the emotion will be recognised and classified in accordance with the categories of ordinary language. The ability to discriminate and recognise complex internal states such as anger, for example, may have to be learnt, and may involve perceptual processes no less complex than recognising a face or a typewriter. (It is possible that the system of classification of internal states is different in different cultures, just like the system of classification of plants, or types of furniture.) The variety and types of second-order motives (Ms) and emotions capable of being generated will depend on the kind of sophistication of these internal perceptual processes, which may vary considerably from species to species, and even between individuals within a culture.

Emotions, moods and attitudes

There are many different mental states and processes which have features in common with emotions. The boundaries between the different concepts are not very sharp, but some broad distinctions can be made. This is not the place for a full survey, but some of the differences are worth noting, since it is easy to confuse emotions with other states, such as moods and attitudes. For instance, a *mood* may be partly like an emotion in that it involves some kind of global disturbance of ones mental processes. But it need not be the intrusion of specific thoughts, desires and inclinations to act, rather a very general tendency to have thoughts and desires of certain kinds. For instance, when feeling depressed a person will tend to make pessimistic predictions, and not notice things which might otherwise have pleased him. Actions and experiences will not be found as satisfying as they normally are. The reverse is true of happiness: there will be a tendency not to be disturbed by violations of one's motives, to be very pleased at simple successes, to expect the best rather than the worst when making predictions, and so on. Thus, there is a connection: having a certain mood often involves being disposed to have certain emotions rather than others. But it need not actually involve having any emotion, with specific motives Mi, Mn, beliefs, Bi, etc.. The dispositions or tendencies may remain mere potentialities. Thus moods involve emotion-generators. In humans, moods, like emotions, are often connected with physiological changes. Chemicals, like alcohol, can change a mood. In addition, moods may be produced by the same sorts of processes as emotions, for instance, when bad news produces depression.

An attitude is a cluster of related motives and beliefs. Attitudes are also sometimes confused with emotions. It is possible to love, pity, admire, or hate someone without being at all emotional about it. The attitude will be expressed in tendencies to take certain decisions rather than others *when the opportunity arises,* but there need not be any continual disturbance or modification of thoughts and decisions. (Addmitedly this notion of a disturbance lacks precision: the boundary we are discussing isn't sharp.) The person you hate may be out of your mind nearly all the time, though when the opportunity to help him turns up, you will be motivated not to, and conversely you will be motivated to do him harm when opportunities for that arise. But you need

not seek out opportunities. In fact, you may even override your desires when the opportunities to vent your hate arise. However, if you hate someone there will be occasions when emotions will be aroused, for instance if he is seen to be happy, or if opportunities occur to thwart his wishes. Thus an attitude may also be a generator of emotional states, though it will be much more specific than a mood. This may be part of the reason why emotions and attitudes are confused. However, an attitude includes motives and motive-generators which may be longer-lasting than the emotions they produce. A parent's love for his child may generate episodes of pride, anxiety, joy, shame, anger. Yet much of the time the child will be out of the parent's mind. The love does not stop then: the cluster of motives and beliefs continues to exist: it is merely dormant. The attitude remains.

Of course, extreme versions of such attitudes can become emotions, if they are not easily put out of mind, and continually disturb other mental processes. If the hate or love is obsessive, and cannot be put out of mind, then it is an emotion. There need not be a very sharp dividing line between attitudes or moods and emotions. The space of possible mental states and processes is too rich and complex for simple divisions to be useful.

There are many kinds of experiences which can be deep and in some sense moving, and which we may describe as emotions, for lack of a richer, more fine-grained vocabulary: for instance delight in a landscape, reading poetry, hearing music, being absorbed in a film or a problem. These seem to involve processes in which what is currently perceived interacts powerfully with a large number of processes, sometimes physical as well as mental. For instance, listening to music can produce a tendency to move physically in time to the music and also a great deal of mental 'movement': memories, perceptions, ripples of association all under the control of the music. These sorts of processes are *not* explained by the present theory, but they might be accounted for in terms of some aspects of the design of intelligent systems not discussed here, such as the need for subtle forms of integration and synchronisation of many processes in controlling physical movement (Sloman 1981). The synchronisation is needed both within an individual and between individuals engaged in co-operative tasks. Music seems to be able to take control of some such processes.

Could a machine have emotions?

The nearest thing at present to a machine with emotions is perhaps a thrashing operating system, trying too hard to share its resources between different processes and consequently not enabling any of them to get on adequately. However, most of the architectural prerequisites for emotions, discussed in previous sections, are not to be found in present day machines. This does not mean they never will be.

Some readers will object that states which lack the physiological disturbances characteristic of many human emotions, are not 'really' emotions, even if they have the other charactersistics. We do not wish to argue about the semantics of 'emotion'. The important point is that there are certain sorts of processes, which play an important role in many of the states we call emotions, and that they are to be expected in intelligent systems. Moreover, it is not the physiological processes that are of most social importance: when you sympathise with the mother of a child recently killed in a car accident, you are not concerned with her physiological state so much as with her beliefs hopes fears frustrations thoughts about what might have been, desires for punitive action, etc. To argue about whether the corresponding mental processes in Martians or machines should be called 'Real' emotions, is quite pointless: the decision will surely be in part arbitrary, either way, a mere matter of definition. (Though how humans react

when actually confronted with such things may not accord with their armchair definitions.)

This is not an exhaustive analysis of emotions: we have at most identified the main features common to a large and important class of emotions. But the word 'emotion' does not have a precisely delineated meaning: what we call emotions shades into other states, such as moods and attitudes. Nevertheless, it should now be clear why we claim that any intelligent system capable of coping with multiple motives in a complex and partly unpredictable world would have mechanisms capable of generating emotions. This does not of course mean (as suggested in an earlier draft of this paper) that they will actually have emotions: that depends on circumstances, as it does with us. If the processes which characterise emotional states are complex combinations of the processes which we have found to be desirable in systems with multiple motives, limited processing speed, and restricted knowledge of a complex world then no special emotion-producing faculty is needed.

Whether all this can be reproduced in a computing system is not obvious yet. Some of the sub-processes are already to be found in AI programs. Some simple sorts of interrupt mechanisms are to be found in operating systems. However, there does not yet exist a machine which could be said to combine all the components of emotions which we have identified. In particular, goal directed AI programs do not at present include large and inconsistent sets of motives with motive generators and motive comparators. Neither do they have the kind of self-monitoring needed for having experiences. Similarly their perceptual, planning, learning and reasoning powers are vastly inferior to those of quite young children, and many animals.

Nevertheless, if what we have sketched represents necessary design features of an intelligent system with multiple motives, limited processing speeds and incomplete or partly inaccurate knowledge of the world, then that implies that designers of super-intelligent machines will need to know something about emotional states, if they want to make sure their machines are going to be controllable.

Conclusion

Every intelligent ghost must contain a machine. The question is what sort of machine? Our sketch of some aspects of possible computational architectures for intelligent systems, provides a framework for studying a variety of mental states and processes in humans, other animals, and perhaps robots. However, the discussion has been restricted to a computational level which is neutral as to the underlying physical architecture. Very little follows from all this about the nature of the human brain, for example, since most of the processes could occur in different sorts of physical mechanisms, just as the same LISP program can run on different kinds of physical computers. Not all the processes and mechanisms described need occur in all minds: different subsets will suffice for coping with different sets of constraints. And some of the important features of human emotions (and cognitive abilities) may depend on physical aspects of the brain.

An interesting open question is whether the construction of a system with the processes we have sketched is *possible.* So far, nothing has emerged in our analysis of motivation and emotions to rule out the possibility of animals or artefacts meeting the specifications. However, it may turn out that there is no way in which a system with so many mechanisms allowing one process to interrupt or disturb others, could function as sensibly as animals and people do. It may be that much tighter overall organisation is required, if the system is not to thrash about while different motives successively try to gain control. Alternatively, it may be that the system of

priorities and thresholds sketched above suffices in most circumstances, but that it is too easily capable of getting into disturbed states which are harmful to the individual. If so, designers of robots should try to find alternative mechanisms: or restrict themselves to designing robots with simpler sets of motivators, or restrict the circumstances in which robots are used.

Perhaps different subsets of the mechanisms and processes sketched above are to be found in different animals. For instance, there may be some animals in which the process of selection of a motive for action (the formation of an intention) is inseparable from the process of initiating action based on that motive. So there would be no operative but dormant motives in such an animal. (This is a less flexible type of system: for example, it cannot interleave a collection of long term plans.) In some animals various types of interruptions may be impossible. Some animals may be capable of selecting intentions prior to execution, but incapable of interleaving their execution. Some may be incapable of storing reasons for actions and therefore abandoning them when goals are accidentally achieved. These and other variations need to be studied in the framework of a generative taxonomy for possible types of mental mechanisms.

The possibility of different combinations of the submechanisms makes it impossible to defend rigorously a strong thesis that that any intelligent system with a complex set of motives and limited powers in a partly unpredictable and partly uncontrollable world will definitely have emotions. Nevertheless, we can see that the possibility of emotions is grounded in mechanisms which seem to be needed for coping with all the constraints we have discussed. A machine would not have emotions if none of its motives was ever insistent enough or had enough merit to interrupt or modify processes generated by other motives. Such a system would have a very *cool* personality. However, it would not be able to respond to sudden dangers and opportunities by rapid interruption and re-organisation of on-going actions. Perhaps there are some people like this? The mechanisms *capable of producing emotional states need not actually do so.*

The analysis of mental concepts presented here has implications for philosophical discussions of the nature of mind and the relation between mind and body. Details will not be discussed here, save to point out that an issue discussed by Anscombe (1957), and several other philosophers can be resolved. The issue is whether X, in 'A wants X' can range over all describable objects or states of affairs. Anscombe claims that the primitive sign of wanting is 'trying to get', whereas we have shown that a want (or other motive) can be present in the system yet never produce any action, or attempted action because it is never selected as operative, for any one of a number of different reasons. (The computational architecture of a less sophisticated computational system than an adult human being, such as perhaps an infant or some other animal, need not support this notion of an inoperative desire.) In addition to the conceptual link with 'trying to get' Anscombe claims that X, the thing desired, must satisfy other conditions, which rule out the possibility of some things being *just* desired for their own sake. Roughly, she claims that in order to want X, A must think there's something good about X. The subtleties of her thesis need not concern us here. Our main point is that on the above analysis, if a representation of X exists in the store of motives, and functions as a motive, i.e. it is able to play the usual role in the process of selecting a motive for action, and in the process of generating action, then that implies that acquiring X is one of A's motives, *no matter how this came to be so and no matter whether A has any justification for wanting X.*

It may be that in a really intelligent system the processes by which motives are added to the store will filter out goals which fail on some additional general requirement of the sort Anscombe describes. Alternatively, the filtering might occur during the process of selecting a subset of motives to become operative, or at some other stage. In that case, the existence of

bizarre motives totally unrelated to the biological or social needs of the organism would indeed be pathological: it would have to result from the break down of the normal mechanisms. But there is no *logical* necessity that there be any such filtering, and it is an empirical question whether it is a normal function of the human mind. It is not part of the ordinary definition of 'want', as Anscombe suggests.

There are still many gaps in the analysis, not least the lack of an adequate computational analysis of experiences of pleasure and pain. See Dennett (1979) for an argument that our ordinary concept of pain is internally incoherent. The notion of desire has not yet been adequately analysed. Neither can we give an analysis of what it is to find something funny. (This is not the same as being able to discriminate things people find funny from those they don't. Finding things funny, like weeping, probably has something to do with being a social animal and the presence of mechanisms whose function is to *force* us to communicate our internal states to others, whether we wish to or not.) It also remains to be shown in detail how complex parallel processes of perception, planning and execution may be capable of being interrupted and redirected in a coherent fashion.

The ideas reported here were developed within the framework of a philosophy of science sketched in chapter 2 of Sloman (1978), which analysed the aims of science primarily in terms of a study of what is possible and an attempt to explain how it is possible. (Laws are merely the limitations of certain ranges of possibilities: the discovery, representation, and explanation of possibilities is more fundamental than the discovery of laws.) The mechanisms postulated here are part of an attempt to explain how a range of types of organisms (and robots), with different kinds of sophistication, might be possible. A full development of that explanation would require analysis of different ways in which the submechanisms could in principle be combined. This should be supplemented with empirical research on the ways they have actually been combined in various organisms.

Finally, the model sketched here may be important for psychotherapy and education. In a system as complex as this, there is enormous scope for 'bugs', some of them already mentioned. For instance, recursive escalation of emotions (the existence of one emotion produces another as a reaction to it, and so on) might account for some catatonic states, not unlike a computer stuck in an infinite loop. Moreover, there are many ways in which the processes by which motives are generated, compared, selected for action, related to planning, triggered when dormant, assigned priorities, etc. may go wrong. The theory sketched here also implies that processes of learning and cognitive development, which are often studied as if they were autonomous, will occur within the framework of a complex and frequently changing collection of motives and motive-generators, capable of disturbing other cognitive processes. These motivators, and the emotional and other processes they generate must have a profound influence on what is learnt when, and it is to be expected that there will be enormous variation between individuals on account of what we previously referred to as the inevitable instability of such a complex system. A deep understanding of the full range of computational possibilities is surely a prerequisite for any systematic attempt to deal with the human problems.

ACKNOWLEDGEMENTS

This paper was written by the first author, based in part on ideas of the second author, who is working on a thesis on emotions, including a critical survey, from a computational viewpoint, of psychological theories of emotions. She has not approved everything herein and cannot be held responsible for errors. Margaret Boden (see Boden 1972, chapters 5 to 7), Jon Cunningham, Arthur Merin, Keith Oatley, Jose Taylor, Stephanie Thornton, and anonymous critics of an earlier version have also helped. A summary of this paper has appeared in *Proceedings International Joint Conference on Artificial Intelligence,* Vancouver 1981. After a draft of this paper had been prepared, our attention was drawn to Doyle(1980), some of which is very similar in spirit. The analysis in H.A. Simon (1967) is also very closely related to ours, though less detailed. Current work by Bob Abelson, George Kiss, Ira Roseman, and Bob Wilensky overlaps with ours. Some of our ideas are a result of collaboration in a project on distributed computing, with Keith Baker, Paul Bennett, Jim Hunter, David Owen, and Allan Ramsay, funded by the UK Science and Engineering Research Council, on Grant GR/A/8067.9. Alison Mudd and Judith Dennison helped with production.

Biblography

Abelson, R.A. in *Proceedings Cognitive Science Conference,* Berkely, 1981.

Anscombe, G.E.M., *Intention,* Basil Blackwell, 1957.

Boden, Margaret *Purposive Explanation in Psychology* Harvard University Press 1972, Harvester Press 1978.

Boden, Margaret *Artificial Intelligence and Natural Man,* Harvester Press, 1978.

Dennett, D.C., *Brainstorms,* Harvester Press, 1979.

Doyle, Jon, *A Model for Deliberation ACtion and Introspection,* Phd.D. Thesis, AI-TR.581, MIT Artificial Intelligence Laboratory, 1980

Dreyfus, H.L. *What Computers Can't Do,* Harper and Row, revised edition 1979.

Dyer, Michael G., 'The role of affect in narratives' unpublished draft, Computer Science Dept, Yale University, 1981.

Dyer, Michael G., 'The role of TAUs in narratives', *Proceedings Cognitive Science Conference,* Berkeley, 1981.

Hayes, P.J., 'The Naive Physics Manifesto', in Michie 1979.

Heider, Fritz, *The Psychology of Interpersonal Relations,* Wiley 1958.

Lorenz, Konrad, *Behind the Mirror,* Methuen, 1977.

McCarthy, J., 'First order theories of individual concepts and propositions', in Michie 1979.

Michie, D. (ed) *Expert Systems in the Microelectronic Age,* Edinburgh University Press, 1979.

Miller, G, E. Galanter and K. Pribram *Plans and the Structure of Behaviour,* Holt, Rhinehart and Winston, 1960

Roseman, Ira, 'Cognitive aspects of emotion and emotional behaviour', presented to 87th Annual Convention of the American Psychological Association, 1979.

Ryle, Gilbert, *The Concept of Mind,* Hutchinson 1949.

Shaffer, L.H., 'Performances of Chopin, Bach and Bartok: Studies in Motor Programming', in *Cognitive Psychology* 13, 1981.

Simon, H.A., 'Motivational and Emotional Controls of Cognition' 1967, reprinted in *Models of Thought,* Yale University Press, 1979.

Sloman, Aaron 'How to derive "Better" from "Is"', *American Philosophical Quarterly,* 1965

Sloman, Aaron 'Physicalism and the bogey of determinism', in S.C. Brown (ed) *Philosophy of Psychology, Macmillan, 1974*

Sloman, Aaron *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind,* Harvester Press, 1978. (http://www.cs.bham.ac.uk/research/cogaff/crp/)

Sloman, Aaron 'Skills Learning and Parallelism', Proceedings Cognitive Science Conference, Berkeley 1981.

Sloman, Aaron and Monica Croucher 'Why robots will have emotions', in *Proceedings 7th International Joint Conference on A.I.* Vancouver 1981.

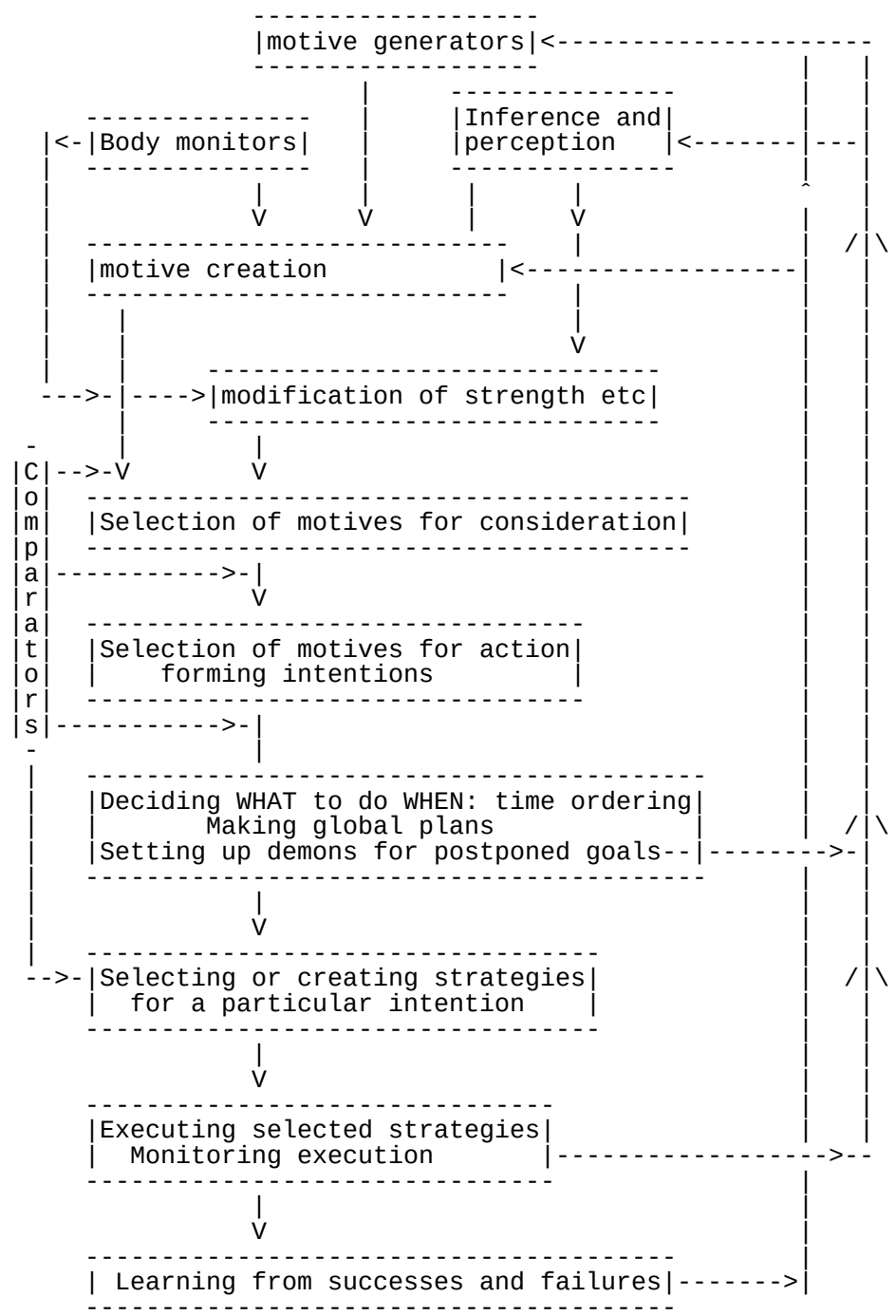
G J Sussman, *A Computational Model of Skill Acquisition,* American Elsevier, 1975

Winston, P.H. *Artifi cial Intelligence,* Addison Wesley, 1978.

Wittgenstein, Ludwig, *Philosophical Investigations,* Basil Blackwell, 1954.

APPENDIX

A schematic outline of some of the main types of processes discussed.



The arrows represent logical order (information flow), rather than time order. It is possible that the different processes occur in parallel. We do not claim that all these processes occur separately in all intelligent systems. But we have argued that separating them and allowing them to occur in parallel improves generality and flexibility, especially since this allows some of the processes to interrupt others, to deal with urgent or important problems. Some of the routes can be ‘short circuited’, for the sake of speed.