APPENDIX TO JCI PROPOSAL
THE "ATTENTION AND AFFECT" PROJECT
Aaron Sloman and Glyn Humphreys

Introduction

This set of notes provides background information for our proposal. The general approach and the detailed concepts and theories are complex and quite difficult to understand quickly because there are so many different ideas involved, from psychology, AI, software engineering, and philosophy, with complex interconnections between them. Moreover they are relatively unfamiliar to most people working in Cognitive Science and HCI. It is not possible to provide a full explanation within the confines of a grant proposal, but it is hoped that this appendix will provide helpful elaboration.

The primary aim is to understand some of the high level functional requirements for the control architecture of intelligent agents with resource limits (primarily limited processing speeds, and limited amounts of parallelism at high levels), embedded in complex and relatively unpredictable environments. On the basis of analysis of these requirements we wish both to explore possible designs and to devise working implementations so that the design ideas can be tested, demonstrated, and exposed to constructive criticism. Any actual implementations will inevitably suffer from a number of simplifications because of limitations of the current state of the art and available resources. Nevertheless, we hope to demonstrate some general design principles and the requirements they fulfil, in a form that survives future detailed elaboration. The most important novel feature of the analysis is the relationship between motive-processing and attention.

The importance of architecture

One of the key ideas, which is fairly obvious but has profound implications, is that the capabilities of a system will depend on its architecture. Moreover, much of the architecture may be specifiable not in terms of physical components, but in terms of "virtual machines" implemented in the physical components, and many of the important capabilities will be definable only in terms of the behaviour of sub-components within the architecture, not directly in terms of input-output relations for the whole system. These are, by now, commonplace characteristics of computing systems, but many philosophers and psychologists have failed to consider their possible relevance to describing and explaining human mental capabilities. However, growing numbers of cognitive scientists have been following the "design-based" approach to understanding mental processes, developed in AI, though not all. Many of those who study affective states still appear to be unaware of the new approach, and very few of those who use the approach have attempted to apply it to the study of affective states, such as moods, emotions, attitudes, and so on. We hope to relate these states to the behaviour of mechanisms concerned with high level control of attention.

From the design standpoint one would not expect to be able to define or explain important states of a complex system on the basis of how it looks to an observer (including itself). Instead, we'd expect to have to explain many of the global states and properties as manifestations of interactions between mechanisms that are intelligible only with reference to the architecture of the system and the functional requirements met by that architecture. To illustrate this consider the difference between describing a system as "slowing down" and describing it as "braking". The former does simply describe a global, externally observable property of the system, whereas the latter presupposes at least an architectural sub-division between a component that has certain causal powers that enable it to vary the speed of operation of some other aspect of the whole system. Similarly, many ordinary descriptions of human mental states, e.g. "believing", "desiring", "intending", "imagining", "angry", "excited", "joyful", "depressed", etc., presuppose complex causal interactions between distinct components of a rich and complex architecture that we are nowhere

near understanding fully. Nevertheless, some global features of the architecture can be described, at least as plausible conjectures worth exploring further.

The space of possible designs

A key idea is that there is a space of possible designs, and actual organisms and machines will have different properties depending on where they are located within this design space. Human beings and other animals are instances of specific but abstract designs that form a small subset of the space. This does not imply either that there is a designer or that there is a unique abstract design corresponding to every individual: the same individual may instantiate a number of different designs at different levels of abstraction. Moreover members of different species may be instances of certain common abstract designs (e.g. being four legged, or having five modes of sensory perception), and members of the same species may instantiate slightly different specific designs. So the space of possible designs is not a simple one-level set: it has a more complex topology, like the multiple inheritance systems in object oriented programming languages.

In order to have a full understanding of the capabilities and limits of a particular organism, it is useful to see where its design features are located within the space, including how they differ from neighbouring designs and what the implications of those differences are. Locating organisms in design space is also relevant to explaining how a particular set of capabilities might have evolved, for instance the ability to have two or more goals at once. Our project addresses a subset of design issues concerned with the relationships between cognitive and affective states in human-like systems.

Architectural design features

Important aspects of a design include (a) the functional requirements that the design satisfies (b) the architecture specified by the design i.e. the number, variety, relationships and functional differentiation of components (c) the kinds of hardware and software components or modules used in the design, i.e. how the design is i*mplemented*. These three levels of description can be applied recursively to components of a design. This paper is mostly about functional requirements.

The different functional roles performed by components within an architecture are determined by their causal relationships to other components. These causal relationships may be of many different types, requiring different kinds of formal models. Control engineers often use collections of partial differential equations to represent causal relations. The causal relations between components of a neural net are not very different from this, though the "emergent" relations between subnets may require very different analyses. At a low level a computer can be described as a mechanism for transforming large bit vectors through a high-dimensional space. Yet another class of causal relations holds between software components in a typical computing system, where procedures may call other procedures and information is passed in the form of parameters and results. Where the programs produce changes in enduring structures and databases, the functional roles are different again. Yet more variety comes from systems supporting not only hierarchically invoked procedures, but also concurrent but asynchronous processes that are capable of interrupting, aborting or modifying one another. Some causal interactions are probabilistic others deterministic. Some involve flow of quantitative information, others structured hierarchically composed data, and yet others simple control signals.

From all this it is clear that the variety of archiectures relevant to understanding intelligent systems is potentially huge. It is very likely that new conceptions of causation and control will be needed for a full understanding of the complex relationships found in human motivation and action.

We shall not be concerned, for now, with low-level differences at the component level (e.g. silicon vs neurones), but only relatively global architectural design differences, mainly at the level of function rather than at the level of implementation. In particular, we'll consider how the requirements of different degrees of sophistication in processing motives can be met by different

architectures. For example, some architectures make it possible to pursue only one goal at a time, whereas others permit multiple goals to be represented simultaneously and acted on with interleaved and combined plans.

Different designs will be capable of supporting different kinds of cognitive states and processes. In particular the ability to have richly structured beliefs and the ability to perform inferences on them imply constraints on possible representational mechanisms (some of which are the result of cultural, some of biological, evolution). Similarly many desires and affective states require equally sophisticated representational capabilities, e.g. hoping desperately that your manager will handle the firing of a dishonest colleague in such a way as to avoid repercussions for you.

This project is not directly concerned with representational issues (though that is one of our long term concerns). Instead it focuses on control issues. One of our conjectures is that mechanisms that satisfy general requirements for resource-limited (especially speed limited) intelligent agents in a complex and changing world will necessarily be capable of producing states that have the essential features of certain kinds of emotional states. In other words there is no need to postulate any special emotional subsystem to account for the existence of emotions. (This is not to deny that there may be some specific emotions, and other affective states, that are produced by specific mechanisms). The project will attempt to refine this conjecture by spelling out design requirements for intelligent agents, and exploring their implications, both by theoretical analysis and by computational experiments. It is not the objective of this project to engage in empirical testing, though because of our links with the School of Psychology, and the interest already shown by other members of the School, the project may in fact inspire experimental investigations, and we shall attempt to spell out testable hypotheses and open empirical questions, in order to encourage this.

Richer architectures imply more varied states and processes

Architectural richness permits increased functional differentiation between states. Diversity of mental states (and processes) requires appropriate architectural diversity. For example, only in terms of details of the information processing architectures, and the different functional roles and causal powers of component mechanisms can we explain the differences between:

- believing that something is the case *vs* desiring it to be the case
- desiring to do something *vs* intending to do it
- intention in current action *vs* intention to do X in the future.
- belief *vs* supposition
- conscious *vs* unconscious processes (if there is any clear difference!)
- hierarchies of dispositions (personality traits, attitudes, preferences, desires, etc.)
- emotional states *vs* various kinds of motivational states involving success, failure, disturbance of plans or actions, etc.

We shall not attempt to analyse all of these cases. Instead we'll try to show, below, how increasing architectural complexity is required for increasing sophistication in processing motives.

There is no "right" architecture

In view of the diversity of forms of animal intelligence it would be foolish to claim that there is some essential kind of intelligence that uniquely determines an architecture. Artificial intelligences can be expected to show even more variation. Corresponding to different kinds of intelligence there will be different architectures supporting different mixtures of mind-like capabilities.

Human mental states require *very* rich and complex architectures, that change during individual development. We have yet to understand what the functional requirements are for human-like intelligence, nor which kinds of designs are capable of fulfilling those requirements. This project will address only a subset of the issues, concerned with attention and the processing of motives.

A more general theory would need to account for a far wider collection of states and processes, including: believing, desiring, learning, perceiving, thinking, supposing, planning, inferring, feeling (sensory, affective), understanding, communication, being puzzled, noticing, attending, acting, finding, finding something funny, aesthetic enjoyment, emotions, attitudes, moods; being aware, conscious, self-conscious, having experiences.

However, the development of a good theory would have the effect of gradually transforming or replacing these concepts must as the development of theories about the structure of matter transformed our concepts of kinds of "stuff", including enabling us to think about kinds of stuff that do not occur naturally. Our existing concepts of kinds of mental states and processes (like ancient concepts of kinds of stuff) are not yet good enough to pose good questions for cognitive science: - they use conflicting criteria, have fuzzy boundaries and lack the kind of generative power for dealing with all the kinds of cases can occur in infants, children, brain-damaged adults, other animals and artificial systems.Many people think they can define mental concepts 'ostensively', but this is an illusion: we don't have reliable self-perception! Instead by studying what is and is not possible for different architectures we can hope gradually to replace muddled ordinary language concepts with new, clearer, more precise ones, related to capabilities supported by different sorts of designs. However, this investigation is still in its infancy.
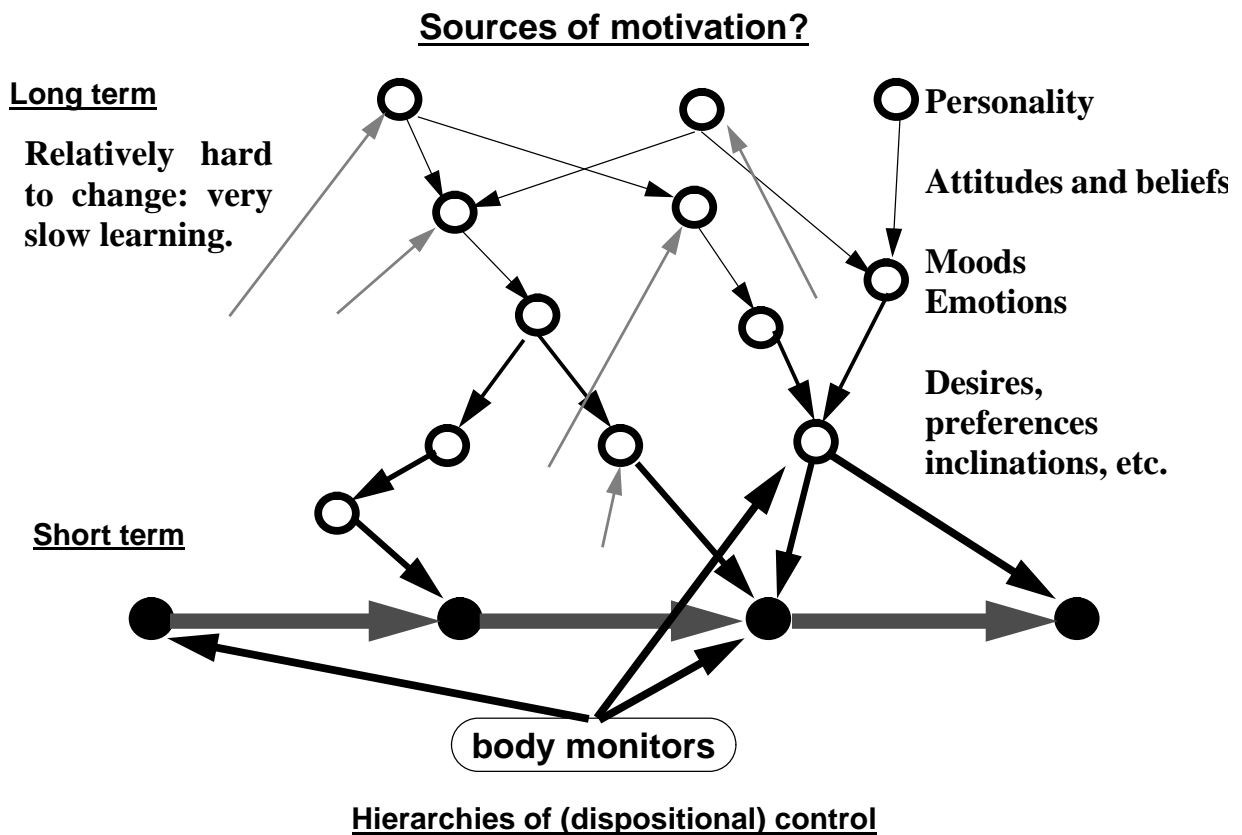
Sources of motivation

Although "intelligence" is not easy to define, it seems that one feature of intelligence in humans and other animals concerns the ability to cope with a complex, partly unpredictable, environment, including other intelligent agents. Another aspect is having a variety of sources of positive and negative motivation ranging from changing bodily needs that generate specific goals, cognitively generated or modified biological desires and dislikes relating to food, drink, sex, satisfying curiosity, through various forms of individual or social activity like games, or parties, to ethical, aesthetic and political ideals and preferences. Some of these are largely innately determined (though shaped by experience) whereas others are mostly a product of culture, and some of them, including addictions and personal taste are specific to particular individuals because of their idiosyncratic history. Some are long term, whereas others are evanescent states triggered by bodily changes or external events, e.g. wanting to help someone who has fallen over. A full discussion of these issues would point to hierarchies of needs and hierarchies of control mechanisms. Some of these mechanisms, operating over long time scales, would be related to personality and character traits. Others, involving shorter time scales, correspond to moods and attitudes. (The word "mood" can be used both to talk about global dispositional states such as "an optimistic mood", "a depressed mood", and about more focussed attitudes and desires as in "I'm in the mood for a walk", "I'm in the mood to go home".) Desires and emotions relate to mechanisms generating states and processes that are still more transitory and more direct in their causal influences.

Human control systems seem to include (several?) hierarchies of dispositions, some very long term and hard to change (e.g. personality, attitudes), others more episodic and transient (e.g. desires, beliefs, intentions).

All involve complex, richly-structured, interacting sub-states, combining information stores and causal interactions that are both context-sensitive (dispositional) and in some cases probabilistic (propensities, tendencies). The following diagram indicates (albeit very vaguely) a possible view of some functional aspects of this sort of architecture, where the circles represent long term and short term dispositions, the black arrows indicate dispositional causal influences of varying strengths, the upward dashed arrows represent the effects of learning on the dispositions, and the shaded arrows represent temporal succession of events (the black blobs) in which particular new motives are generated or modified. An example of a long term disposition might be a tendency towards benevolence. This could produce a more specific and more easily changed attitude of

benevolence towards a particular group of people. This in turn, might interact with particular episodes to produce very strong friendship towards an individual member. Knowledge of impending danger to that person might produce a particular, short-lived desire to act, and so on. If this desire has a strong tendency to gain control of attention (high "insistence") then the state can be described as an emotion, in which the agent has some loss of control.

Although the diagram suffers from considerable ambiguity and vagueness, it is clear that fhere are many different ways this sort of architecture could be implemented. For example the intermediate dispositional states could be implemented in neural nets whose weights are influenced by higher level nets, or production systems whose rules are modified by higher level systems, or.... There are many different mechanisms that might implement such architectures. **What** they do is more important than **how** they do it.

.

Sources of motivation?

Long term

Relatively hard to change: very slow learning.

Personality

Attitudes and beliefs

Moods Emotions

Desires, preferences inclinations, etc.

Short term

body monitors

Hierarchies of (dispositional) control

So far nothing has been said about what happens to new motives. It is instructive to consider a succession of increasingly complex kinds of cases, with increasingly sophisticated architectural requirements (which are orthogonal to the requirements implied by the above diagram). Only when we get to the fifth case will the need for mechanisms that deal with interactions between new motives (or other new information) become clear.

State transitions for motives: Case 1

Different architectures can support different kinds of sophistication in the processing of motives. A very simple architecture would support only one goal at a time with few processing options, as indicated in the following diagram. NB. the diagram does not represent an architecture. The nodes and arrows represent transitions through which a particular motive can pass. These are not

necessarily states of *the whole system,* as we shall see later, when we consider systems that can have several motives passing through different states.
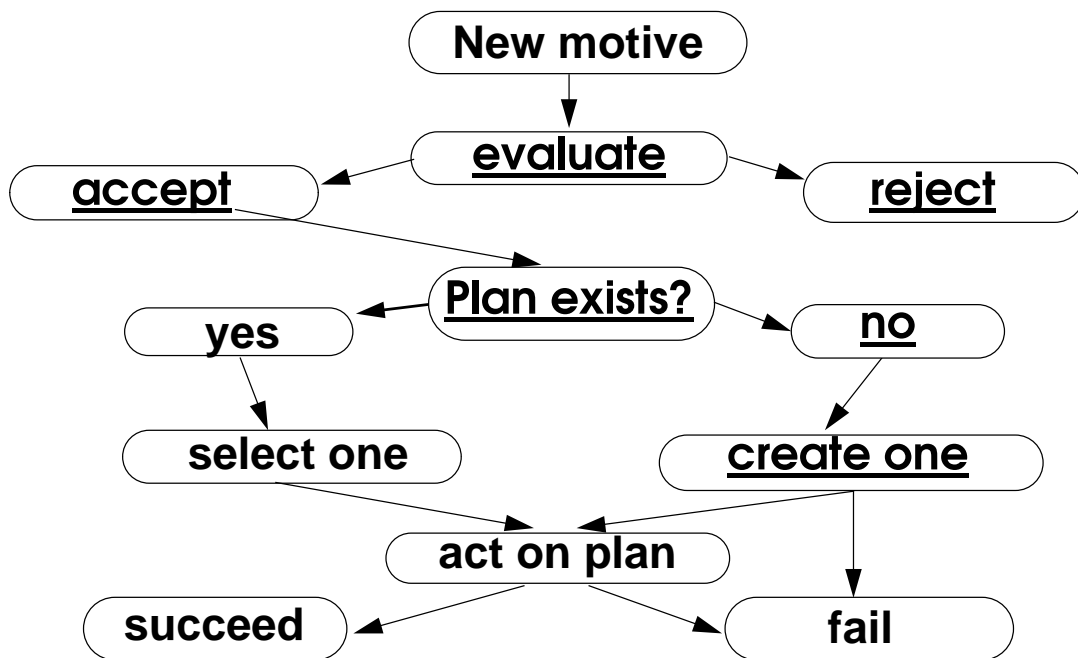
```
        ┌─────────────┐
        │ New motive  │      Compare: a command-
        └─────────────┘      driven software package.
               │
               ▼
        ┌─────────────┐
        │ Select plan │
        └─────────────┘
               │
               ▼
        ┌─────────────┐
        │ Act on plan │
        └─────────────┘
          ↙         ↘
┌──────────┐         ┌──────────┐
│ succeed  │         │   fail   │
└──────────┘         └──────────┘
```

Case 1: An agent with stored plans and only one motive at a time

So the diagrams representat *transitions of substates* of the whole system. Each diagram will appear to be capable of being implemented on a simple finite state automaton. We'll later see the need for something more complex.
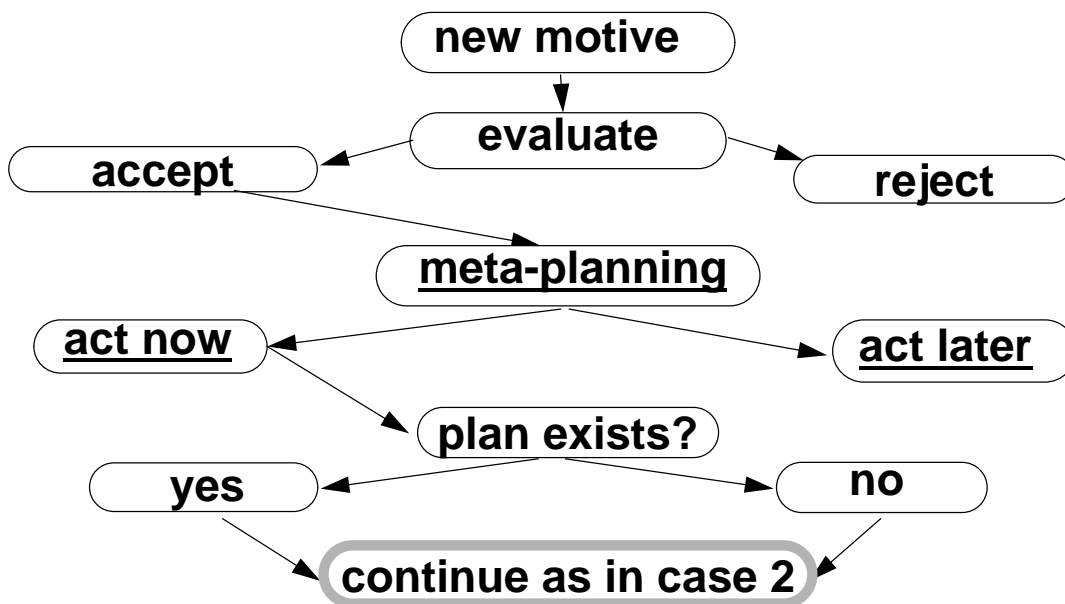
State transitions for motives: Case 2

```
              ┌─────────────┐
              │ New motive  │
              └─────────────┘
                     │
                     ▼
┌──────────┐   ┌──────────┐   ┌──────────┐
│  accept  │◄──│ evaluate │──►│  reject  │
└──────────┘   └──────────┘   └──────────┘
        ╲
         ╲    ┌───────────────┐
          ──►│  Plan exists?  │
             └───────────────┘
              ↙            ↘
      ┌──────────┐        ┌──────────┐
      │   yes    │        │    no    │
      └──────────┘        └──────────┘
            │                   │
            ▼                   ▼
   ┌──────────────┐     ┌──────────────┐
   │ select one   │     │  create one  │
   └──────────────┘     └──────────────┘
             ╲          ╱          │
              ╲        ╱           │
           ┌──────────────┐        │
           │ act on plan  │        │
           └──────────────┘        │
             ↙          ↘          ▼
    ┌──────────┐      ┌──────────┐
    │ succeed  │      │   fail   │
    └──────────┘      └──────────┘
```

Case 2: Some agents can evaluate and reject motives
and create plans if necessary

The second diagram indicates motive state transitions for a mechanism that can still handle only one motive at a time, but which evaluates it and decides whether to act on it or not, before

attempting to find or create a plan. Moreover, unlike the previous case, this system can also create new plans if necessary.The additional possible states are indicated by underlined labels, some of which hide considerable complexity! It is clear that a mechanism with these capabilities would require a more complex architecture than the previous system with additional sub-mechanisms required (a) to support the evaluation process and (b) to support the creation of plans if an appropriate one does not already exist.

State transitions for motives: Case 3

The ability to postpone attending to a new motive requires additional architectural complexity. The next diagram illustrates state transitions possible when a meta-planning submechanism is added, i.e. something that can determine whether a new motive should be acted on immediately, or action postponed. As before underlining indicates new capabilities.
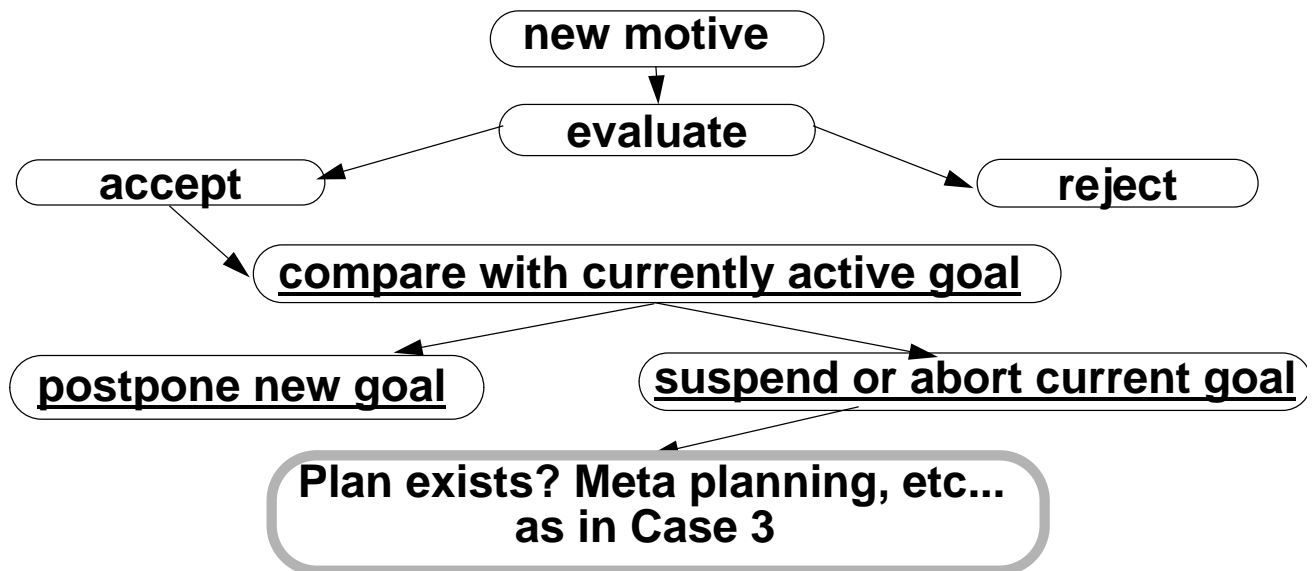

Case 3: State transitions for a system able to postpone motives

In fact the situation is bound to be more complex than the diagram suggests, because if acting on a new motive is to be postponed then an additional store is required for postponed motives, and additional sub-mechanisms are needed for deciding that the time has come for the postponed motive to be acted on. Which mechanisms are needed will depend on what the conditions for reactivation are, and whether the reactivation can occur while another motive is currently being acted on.

State transitions for motives: Case 4

An intelligent agent should be able to work out whether a new goal is more important and/or more urgent than something currently being pursued. This requires additional states to be added to the motive transition graph, and additional architectural complexity that allows the comparison between new and old goals to be done in the midst of executing the old goal. In some cases this can be done by temporarily suspending the current action. Human beings do not always have to do that: sometimes they can evaluate a new goal (e.g. whether to help someone who appears to have fallen over on the other side of the road) while stlil acting on a previous motive (e.g. walking to catch a bus). Some of the additional complexity of processing is indicated in the next diagram,

though as before it does not provide a precise or complete specification. In particular, it does not distinguish cases where the original activity can be continued after the old one has been completed, from cases where the only options are to abandon the old action completely or to postpone acting on the new motive. Different architectural additions are required to support the different kinds of extra functionality.

```
                          ┌─────────────┐
                          │ new motive  │
                          └─────────────┘
                                 │
                                 ▼
                          ┌─────────────┐
            ┌─────────────│  evaluate   │──────────────┐
            │             └─────────────┘              │
            ▼                                           ▼
   ┌─────────────┐                            ┌─────────────┐
   │   accept    │                            │   reject    │
   └─────────────┘                            └─────────────┘
            │
            ▼
   ┌──────────────────────────────────────┐
   │  compare with currently active goal   │
   └──────────────────────────────────────┘
            │                    │
            ▼                    ▼
┌─────────────────────┐  ┌──────────────────────────────┐
│ postpone new goal   │  │ suspend or abort current goal │
└─────────────────────┘  └──────────────────────────────┘
                                     │
                                     ▼
                    ┌────────────────────────────────────┐
                    │ Plan exists? Meta planning, etc...  │
                    │            as in Case 3             │
                    └────────────────────────────────────┘
```

Case 4: Additional transitions possible if current action
can be interrupted, aborted, etc

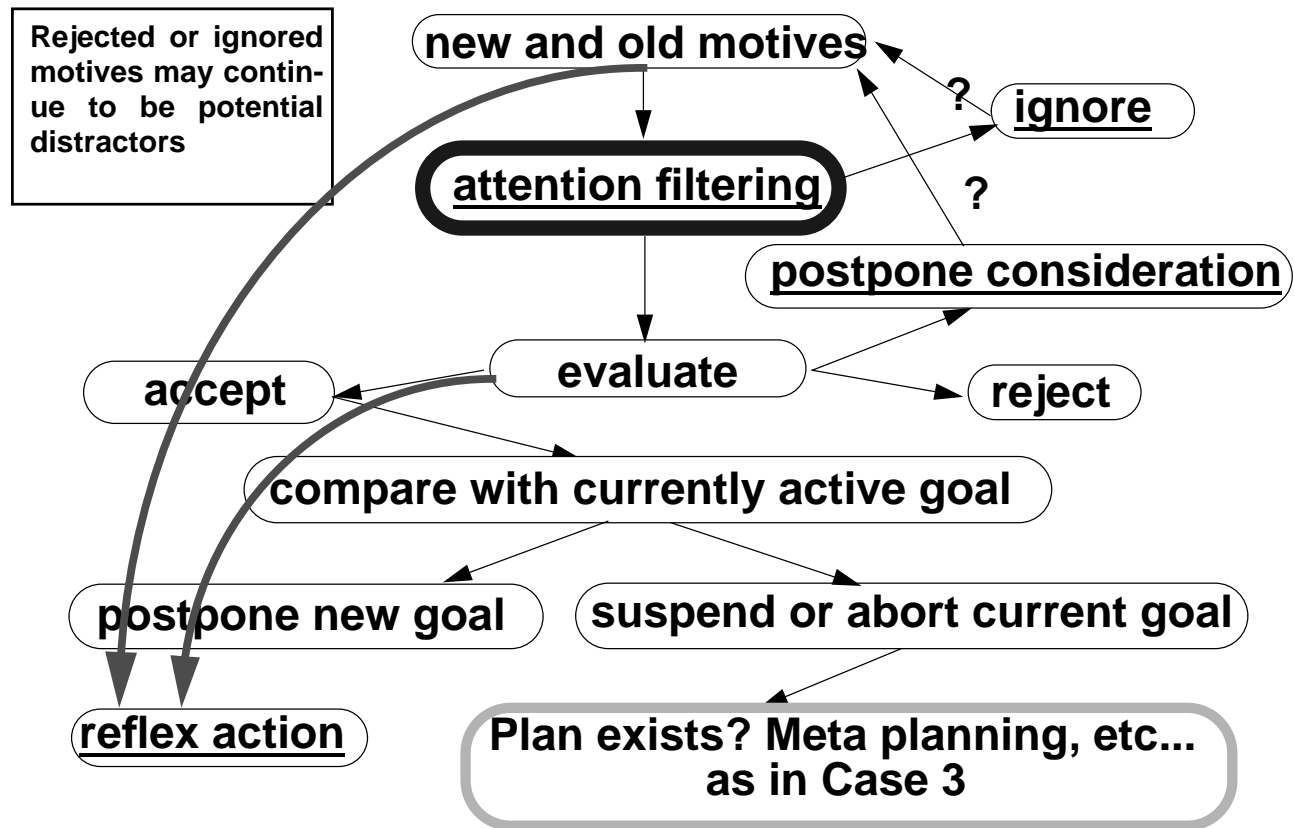## State transitions for motives: Case 5

Evaluation of new information can consume time and other resources (e.g. the eyes may need to be diverted to check something), and disrupt important activities. One way to deal with this is to spend only enough time investigating a new motive to decide whether it is worth investigating further by comparing it with currently active processes, etc. If not, further consideration can be postponed till later. However, there are situations where even this shallow consideration would interfere with some very important ongoing action in a drastic way e.g. a brain surgeon in the middle of a particularly delicate operation, someone being given important instructions verbally, where getting something wrong later could be a matter of life or death, or a motor-cycle racer going round a bend at high speed, with competitors close by.

It is not at all obvious why a highly parallel system like a human brain should have such limitations in what it can do in parallel, and that is one of the design issues we have yet to understand. It could be a consequence of a hierarchic control structure which is desirable for other reasons, or it might be a side-effect of a particular implementation of high level control, for instance using a competitive neural net. The answer will probably be found after close examination of the differences between those cases where we can attend to two things at once and cases where we cannot. There is clearly some individual variability: some people freeze in certain tasks (e.g. washing up, or eating) if spoken to, whereas others do not. The authors of this appendix can work while listening to music, whereas some people find it too distracting.

Nevertheless, there always seem to be only a relatively small number of things with any intellectual content that people can do at once. This seems to be such a significant feature of

human resource limits that it probably follows from some as yet unknown functional requirement for the high-level control mechanisms.

In view of such limited internal processing abilities, and the need to "protect" precious and urgent motives, it is desirable to have a mechanism that somehow prevents attention being diverted in such cases. The next diagram indicates some of the motive state transitions that can occur in a system that has attention filtering capabilities.

Rejected or ignored motives may continue to be potential distractors

new and old motives

ignore ?

attention filtering

? postpone consideration

accept ← evaluate → reject

compare with currently active goal

postpone new goal   suspend or abort current goal

reflex action   Plan exists? Meta planning, etc... as in Case 3

Case 5: Attention filtering, reflex decisions, insistent rejected motives

---

In addition to filtering, the diagram also indicates that some motives will, if they have reached a certain stage of processing, trigger some automatic reaction that happens quickly without any consideration of whether or how it should be done, or creation of a new plan. There seems to be plenty of evidence that human beings can be trained to have new cognitive "reflexes" in situations where there is no time for consideration of whether to act and if so how, etc. Provision for such reflex reactions requires additional architectural complexity, and if there are mechanisms that can _learn_ new automatic reactions, as opposed to having only "hard-wired" reflexes, then even further architectural complexity is required.

Notes on attention filtering

So far we have said nothing about *how* the filtering is to be done: it may or may not make use of an explicit filtering mechanism. Whatever happens, it must be done very quickly and without making use of elaboring reasoning about whether the new information should or should not receive attention, for that would defeat the very purpose of filtering. That means that it will have to use fast (and therefore unreliable) heuristics. It is possible that these heuristics will also be

modifiable by training. People who have learnt that a particular sound is an indication of impending danger may notice it and think about what to do in situations where others would not be distracted by it. It is also necessary that the filtering process be context-sensitive: something that distracts attention normally should not be able to do so when it would interfere seriously with another activity. Learning which activities should not easily be diverted requires additional mechanisms.

It is possible that something that fails to divert attention, such as a sound or a movement, or a motive to deal with a sensation of in insect on the skin, does not persist, and the chance is then lost forever. In other cases whatever produced the motive may continue to do so, so that although it does not actually get through the filter the dispositional state in which it is "attempting" to do so can persist for some time, until either the cause goes away or the filtering conditions change and it gets through. Similarly, even if a motive which gets through is considered briefly and a decision taken to postpone further consideration, it may still continue "attempting" to divert attention, and in some cases may repeatedly succeed, even though an explicit decision has been taken to reject the motive in question. This is an example or a partial loss of (mental) self-control that is characteristic of certain emotional states, including fear, grief, anger, jealousy, excited anticipation, being thrilled at a recent achievement, etc.

Simultaneous processing of multiple motives

Each of the above graphs indicates state-transitions possible for a single motive, though the "*compare with currently active goal*" sub-processes presuppose the coexistence of more than one motive. In the case of an agent with multiple independent sources of motivation, with sub-mechanisms able to generate new motives asynchronously, there could be several coexisting motives all at different stages of processing (including some that have not managed to divert attention, but still persist, and others that have been considered and rejected but still persist). Clearly yet more architectural complexity is required for handling the collections of postponed motives, the actions that may have been temporarily suspended in order to make way for more urgent activities, and so on. Nothing has been said here about the ongoing "monitoring" required to enable a system to detect the conditions under which postponed motives should be re-considered, or suspended actions resumed, and so on. In additon to all this concurrency there is a general need for continual monitoring of the environment. In the case of human beings there appears also to be a continual process something like ruminating on prior knowledge. Thus new motives can arise from remembering something, without any external trigger.

The requirements sketched here seem to require "coarse-grained" concurrency rather than the kind of parallelism investigated in connectionist models, though it is possible that connectionist implementations of coarse-grained concurrency and resource-limited high level mechanismswill be found useful.

Criteria for assessing motives

As new motives are generated, an intelligent system has to decide which to act on, and when, and how. The evaluation process can depend on, for example:

- **Importance** this will be a partial ordering and may have different dimensions. It may use some kind of numerical measure or only descriptive rules ("things of type X are more important than things of type Y")
- **Urgency** In simple cases all that is needed is a measure of how much time is left before it will be too late to act on the motive. In more complex cases (as Luc Beaudoin has pointed out) there will be a time-varying measure of the difficulty of acting on the motive, or the likelihood of success, or the undesirable side-effects. Taking account of the latter is far more complex than simply feeding a measure of urgency into some decision algorithm.

These assessments can be arbitrarily complex, and use arbitrarily sophisticated knowledge and reasoning, which is why it is necessary to have some filtering process that prevents consideration

of these issues, in some circumstances. Both Importance and urgency of a motive should not be confused with:

- **Insistence** This is a measure of ability to distract attention. In order to be useful, it should function as a rapidly computed heuristic measure of the **potential** urgency and importance of a new motive.

More aspects of motive processing

Motives may be postponed till a condition is satisfied, till a specified time, till some unspecified future (i.e. put on a "wish list") etc.

There are different sorts of "reflexes" more or less subject to training, to voluntary control, to monitoring. (Compare fluent reading, musical sight-reading, playing table-tennis, etc.).

Plans may be more or less explicit, more or less complete, more or less subject to voluntary intervention. "Reactive planning" using pre-stored routines invoked by recognizing situations in which they are relevant obviously has a role, especially in connection with reflex responses. "Anytime planning" uses strategies that enable *some* useful result to be provided even if planning has to be terminated prematurely. E.g. it may create a high level overview of what is to be done, or an indication of how to get going, or an analysis of the most difficult stage of the action, leaving other details to be filled in while acting. If the time available for assessing importance and urgency of motives, and for making plans, is known to be limited, then "deliberation scheduling" may take place, providing a basis for deciding what to think about, to how much depth, and in what order.

In some cases two or more activities can proceed in parallel (walking and thinking or talking), in others not (thinking about your holiday while listening carefully to complex instructions).

Not all this richness is to be found in young children: some of the architectural complexity *develops over time.* How?

The filtering and other resource-limited mechanisms that use heuristics and short cuts will be inherently fallible. This reduces the scope for super-intelligent, super-rational machines.

Varying architectural requirements for motivational processing

It should by now be clear that different architectures support different kinds of motivational complexity. It remains an open question how much of the functionality described here is to be found in different animals, and whether similar sub-mechanisms are used for the same purposes or whether very different implementations are used.

A motive can be either a descriptive representation of something to be achieved or done, or a control signal of some kind, or a mixture of some kind, or..... How this list should be continued is a topic for further research.

Exploring the architectural requirements for different organisms and machines requires consideration of at least the following kinds of variation in capabilities and designs:

- Differences in how motives, plans, etc. are represented.
- Differences in planning and learning capabilities.
- Single high level goal vs multiple independent goals
- Single current plan/action vs multiple (possibly interleaved) actions.
- Static priorities vs dynamic revision of priorities.
- Evaluation of goals in isolation vs evaluation of goals relative to other existing goals.
- Monitoring driven explicitly by current plan *vs* parallel asynchronous general-purpose monitoring, *vs* a combination.
- Interrupt capabilities of various kinds. Interrupting action *vs* interrupting/diverting "attention".
- Whether during certain intricate and dangerous tasks, new questions, thoughts, inference processes, can or cannot be processed without serious risk.
- Various more or less explicit mechanisms for "attention filtering".

Some of the differences in mechanisms used by different organisims or machines may be marginal, if the encompassing architectures are functionally equivalent.

Sources of new motivation

Depending on the architectural richness of the system, there can be different sources of motivation, including:

1. **Body monitors:** Direct, or via inferences
2. **New percepts**

- Synchronous (sub-goals related to plan execution)
- Asynchronous E.g. produced by monitors. May use either 'reflex' responses, or reasoning.

3. **Inferences from old beliefs**:

Reflecting on something you know generates a new subgoal for a pre-existing goal.

4. '**Triggering' by thoughts.**

Some motive-generators are innate, some trained reflexes, some rules applied in rational, conscious deliberation.

An intelligent system needs recursively modifiable motive generators, motive comparators, motive generator generators, motive comparator generators, etc. (We can represent these as mg mc mgg mcg mgc mggg mggc, etc.)

Anger - an example of an emotion

What is required for X to be angry with Y for doing Z?

1. X believes P1
2. P1 is the proposition that Y did Z
3. P1 being true conflicts with one of X's desires
4. (1) and (3) cause X to acquire a new desire P2
5. P2 involves X doing something unpleasant to Y
6. X believes that Y had no right to do Z
7. X believes that P1 justifies P2
8. X is disposed to be pleased if unpleasant things happen to Y

NB So far these are not _sufficient_ conditions for X to be angry. Perhaps X simply _wishes_ Y had not done Z, so he coolly and calmly decides that it would be a good thing if Y were made to suffer for this. Anger requires something more: X being "moved", i.e. not being in complete control of his mental processes. This is a subtle dispositional property involving the tendency of P1 and P2 to get through attention filtering., so we have condition:

9. Thoughts, motives, etc. relating to P1 and P2 keep intruding I.e. the new states have high 'insistence' and tend to divert attention (depending on the current filter threshold).

It was this kind of consideration that originally inspired the architectural analysis including a requirement for attention filtering.

There appear to be cases of emotions like anger that satisfy only the conditions 1 to 9. However, there is strong (and futile) disagreement over what "emotion" and "anger" really mean, and some people require additional conditions to be satisfied. Ignoring irrelevant questions about the "real" meaning of terms we can acknowledge that there are some cases where at least two additional conditions seem to be fairly closely bound up with the above states, namely:

10. X is _aware_ of facts 1 to 9 (or some subset thereof). This requires some kind of _self-monitoring_ capability. Otherwise we have a case of _being_ angry without _feeling_ angry. The former does not need self-awareness.

11. The above phenomena may (but need not) cause physiological 'arousal'. Some emotional states include muscular tension, flow of adrenalin, frowning, sweating, etc. and X's experience of

his own state is altered by this. But they are not **essential** for emotions if conditions 1 to 9, or 1 to 10 are considered sufficient for an emotion..

Conjecture:

The architecture needed for resource-limited intelligent agents in a complex, fast-changing, partly unknowable environment, suffices to explain the possibility of certain kinds of emotional states. In particular, mechanisms concerned with the sorts of filtering mentioned in case 5 seem to be required for the existence of dispositional states in which thoughts and motives have a high "insistence", as in condition 9 above for anger.

Warning:

The concept 'emotion' is ill defined. People muddle emotions, attitudes, moods, desires, feelings of various sorts, and even aspects of character and personality. So it is not worth arguing about which sorts of architectures are essential for emotions and whether emotions have a function or are simply combinations of side-effects of other processes with functions, or whether there are specific mechanisms that evolved in order to create emotions, etc. What is clear is that processes of the kinds described in conditions 1 to 9 for anger, and possibly also condition 10, could exist in intelligent agents designed to cope with the functional requirements discussed in connection with limitations of attention, and without a need for any additional mechanism specific to emotional states.