Design Requirements for a Computational Libidinal Economy

Ian Wright
Cognitive Science Research Centre
School of Computer Science
The University of Birmingham
Birmingham, B15 2TT, England
I.P.Wright@cs.bham.ac.uk

July 2, 1996

Abstract

Design requirements for a computational libidinal economy are presented that constitute a preliminary theory of basic types of motivation and learning. The theory avoids many of the difficulties of Freudian libido theory and has new arguments in favour of it. A corollary is a circulation of value theory of simple affect that builds upon existing information processing theories of emotion. Such a theory can account for some forms of cognitive pleasure and unpleasure, in particular the feelings involved in attachment and loss.

INTRODUCTION

Some emotional episodes, such as grief and loss, characteristically involve states of intense 'mental pain', a form of unpleasure not linked to a part of the body in organic distress. From the first-person perspective, the 'mental pain' of such emotional episodes is arguably their most prominent feature. For example, mourners often use analogies with physical distress when describing their mental states (Wright, Sloman & Beaudoin, 1996), employing words such as 'hurt', 'pain' and 'intensity'. The aim of this paper is to offer a preliminary account of the functional role of certain forms of mental pleasure and unpleasure and their connection with emotional states.

It is a recurring assertion that there is a relative neglect of motivation and emotion in cognitive science. For example, (Simon, 1967) replies to Neisser's criticisms that information processing theories of mind cannot account for feelings and emotions. More recently, (Newell, 1990) lists motivation and emotion as missing elements that need to be included in more comprehensive information processing theories of mind. Introducing the diversity of phenomena that can be classified as 'motivational' or 'emotional' to unified theories of cognition is no small task, and will not be solved tomorrow. This document, therefore, attempts to explain only a subset of motivational and affective phenomena. The hope is that a narrower explanatory scope may be more theoretically tractable, perhaps affording generalisation to other phenomena later.

The problems of motivation and emotion are considered from the perspective of artificial intelligence, in particular from research involving the design and implementation of agent architectures. A very simple design hypothesis is described that generates a set of high level design requirements for a motivational subsystem. The requirements for this system specify a computational libidinal economy that purports to explain a subset of motivational and emotional phenomena. The theory is new, but has important antecedents in Freudian metapsychology and modern information processing theories of emotion.

Throughout a *design-based* approach is taken; this involves taking the stance of an *engineer* attempting to build a system that exhibits the phenomena to be explained. It involves exploring an abstract space of possible requirements for functioning agents (*niche-space*) and the space of possible designs for such agents (*design-space*) and the mappings between them (Sloman, 1995a). This is an iterative process, and one we have been pursuing in the Cognition and Affect group. The work presented here builds on a continuing research effort detailed elsewhere (e.g., (Beaudoin & Sloman, 1993; Beaudoin, 1994; Sloman, Beaudoin & Wright, 1994; Wright, Sloman & Beaudoin, 1996)). To date all our implementations lag behind theory; however, the process is often more important than the product: attempting to develop a design to meet a set of requirements helps clarify many issues. For example, section 2 summarises work done mainly by Aaron Sloman and Luc Beaudoin. It outlines some requirements for autonomous agency and a partially implemented coarse-grained architecture designed to meet the requirements. Sloman's attention filter penetration theory (*afp*) of emotion, which builds on the work of Simon, is then briefly discussed, and the notion of *perturbance* introduced. Section 3 presents examples of emotional states that can be classified as *valenced perturbances.* The *afp* theory does not account for such states. This motivates an investigation of valency.

In section 4 evidence is presented for the existence of a *quantitative, universal representation of value* in human motivational subsystems. This is the preliminary to section 5 that lists requirements for a computational libidinal economy situated within an agent architecture, and two corollaries: a circulation of value theory of simple affect, and a reappraisal of libido theory. The circulation of value theory is briefly compared with existing information processing theories of emotion, in particular how it can provide a unified account of valenced perturbances, linking together such phenomena as mental pleasure and unpleasure, allocation of processing resources, basic motivators and reinforcement learning. Section 6 concludes with a discussion of the explanatory power and limitations of the computational libidinal economy and thoughts on its possible development and prospects for working implementations.

AN ARCHITECTURE FOR MOTIVE MANAGEMENT

In this section some requirements for human-like agency are outlined briefly, followed by a design for an architecture that attempts to meet those requirements. This kind of architecture can generate emergent processing states, called perturbances, which are characteristic of many emotional states.

Requirements for agency

The requirements for successful operation in dynamic, unpredictable, real time environments pose certain design problems. For example, the current situation (both externally in the environment, and internal within the agent) provides 'too much' information (Hayes-Roth, 1990): the agent, having limited time for processing and selecting, will need to focus its processing and ignore irrelevances. Information will be widely distributed both in time and space, requiring the agent to search for relevant information and remember, recall and integrate past information. Diverse demands will be placed on agent functioning: for example, at one moment the agent may have very little to do and have the luxury of deliberation while at the next moment the agent may need to perform many complex tasks quickly and efficiently. The unpredictability of the environment renders complete planning prior to action impossible. Instead, opportunities and threats to plans will need to be constantly monitored for. An autonomous agent needs to be robust, flexible and have the ability to cope with variable stress (Beer, Chiel & Sterling, 1990).

Actions need to be globally coordinated. This requires the ability to select between multiple motives (Sloman, 1985), prioritise goals, decide on a level of commitment towards current intentions, and notice opportunities for actions that satisfy more than one motive. To react to new, motivationally relevant events in the environment the agent will need to interrupt its ongoing processing and switch its 'attention' to new contingencies (Sloman, 1987).

To detect such events it must be able to generate motivations asynchronously to current processing. A level of coarse-grained parallelism is therefore necessary (Simon, 1967; Sloman, 1978; Maes, 1990) to enable execution of current goals and at the same time check for new information that may entail goal revision. The agent will need to be adaptable and learn from mistakes and successes. Therefore, architectures that model autonomous agency will need to integrate a wide range of behavioural capabilities. (Bates, Loyall & Reilly, 1991) term such architectures 'broad'.

An agent's computational resources are bounded. For example, humans find it difficult to listen to more than one conversation at once. Also, the agent will be physically constrained – it will only be able to move at a certain pace, manipulate a finite number of objects and so on. Good design solutions will manage an agent's finite resources as efficiently as possible, although efficiency is a difficult notion to define (Sloman, 1995b).

The following section describes a 'broad but shallow' architecture that is intended to meet these kinds of requirements. It is 'broad' because it integrates many capabilities, and 'shallow' because many of these capabilities are initially designed and implemented in a simplified manner.

A design for agency

A full account of the proposed architecture [1] would be too long for this paper, so a highly summarised account of the processing of motivators is provided. Figure 1, an impressionistic diagram, shows the coarse functional breakdown of the architecture.

[1] There are are at least two uses of the word 'architecture'; one referring to an abstraction or design that is common to many instances of the architecture; and the other to concrete instances of such designs. The former sense is used here, in which an architecture is a collection of features common to a class of entities. Each instance of an architecture is composed of coexisting, interacting substructures with various capabilities and functional roles. A substructure may also have an architecture. The architecture of a complex system can explain how its capabilities and behaviour arise out of the capabilities, behaviour, relationships and interactions of the components. An architecture can be specified at different levels of detail, for example at a high level of abstraction the architecture of a house will not include the occurrence of particular bricks, whereas a more detailed architectural specification would (Wright, Sloman & Beaudoin, 1996).

The large shaded area represents ‘automatic’ processes (associative memory, low level sensory analysis, low level motor control processes, innate and trained reflexes) all implemented in highly parallel dedicated, but trainable, ‘hardware’. It is assumed that such processes include mechanisms shared with many other animals. The larger unshaded area above represents ‘management’ processes involved in (among other things) deciding whether new motives should be adopted or not. The management processes, but not the low level processes, can create, consider and evaluate explicit representations of options before selecting between them, for deliberation, planning and problem-solving. It is suspected that most other animals do not share this capability with humans: their architectures are not sufficiently rich.

The architecture has many components which coexist and operate concurrently. All the components require a depth and complexity that we cannot describe here. For more details, and for the historical progression of these ideas, see, for example, (Sloman, 1978; Sloman & Croucher, 1981; Sloman, 1987; Beaudoin & Sloman, 1993; Beaudoin, 1994; Wright, Sloman & Beaudoin, 1996). The architecture is a virtual (abstract) machine whose components and processes need not correlate in any simple way with physical structures and processes.

The architecture has *a changing collection of motivators.* A motivator is a semantically rich information structure that tends to produce, modify or select between actions. ‘Motivator’ subsumes desires, goals, intentions and wishes. It typically expresses a motivational attitude (‘make false’, ‘keep true’, etc.) towards a possible state of affairs (‘short of food’, ‘warm’, ‘in danger’, etc.), which may be expressed in propositional or non-propositional form (Beaudoin & Sloman, 1993). Motivators have various associated information items, including urgency, importance and an *insistence* value (Sloman, 1987). The insistence of a motivator is its dispositional ability to gain management processing resources.

Motives are generated by *preattentive and attentive motive generactivators* (Beaudoin, 1994). These express agent ‘concerns’ (Frijda, 1986; Moffat & Frijda, 1995), which can be defined as dispositions to desire occurrence or nonoccurrence of a given kind of situation. They operate asynchronously in parallel, triggered by internal and external events. They

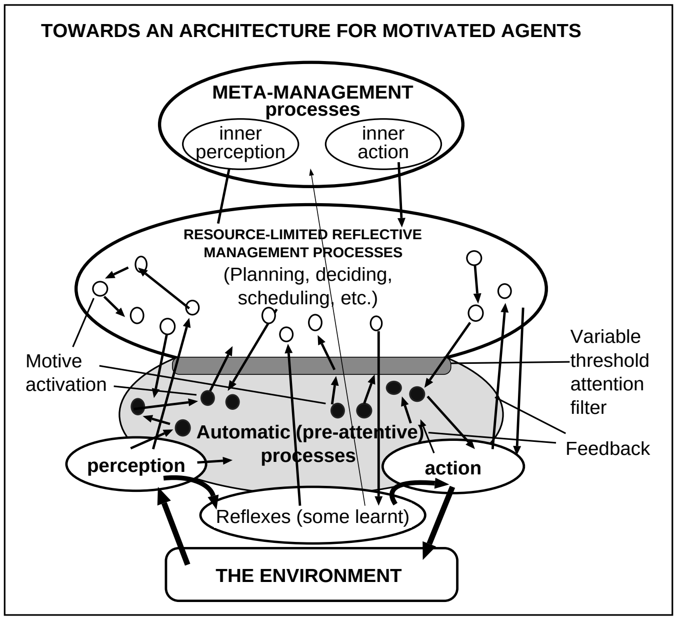
TOWARDS AN ARCHITECTURE FOR MOTIVATED AGENTS
META-MANAGEMENT processes
inner perception
inner action
RESOURCE-LIMITED REFLECTIVE MANAGEMENT PROCESSES
(Planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
Automatic (pre-attentive) processes
Feedback
perception
action
Reflexes (some learnt)
THE ENVIRONMENT

Figure 1. Towards an architecture for motivated agents

generate and activate or reactivate motivators, and set or reset their insistence level. Insistence is a 'quick and dirty' heuristic measure of the *urgency* and *importance* of the motivator. For example, a generactivator to maintain fuel levels may activate a highly insistent 'fuel seeking' motivator.

There are both empirical and design arguments supporting the claim that despite parallelism in the brain high level cognitive processes are resource-limited: not all potentially useful management tasks can be performed concurrently. This 'processing-limit' leads to a requirement for attention filtering, or motive selection. A *variable threshold attention filter* protects resource limited management processes from unnecessary interruption by allowing only items with insistence values 'higher' than the current filter threshold to divert attention. Motivators that fail to surface may remain available, so as to take advantage later of a lower filter threshold, or they may die unless continually reactivated by the relevant generactivators. Motivators survive until they are satisfied or decay.

Motive management is a set of processes that occur once a motivator has ‘surfaced’. A surfaced motivator can trigger many diverse and complex processes (Beaudoin, 1994), including: assessing (evaluating its importance, urgency, costs and benefits etc.), deciding (whether to adopt the motive, i.e. form an intention), scheduling (when and under which conditions to act), expansion (how to do it, i.e. planning), prediction (projecting the effects of hypothetical decisions), detecting motive conflicts, detecting opportunities, abandoning a motive and changing filter thresholds (Sloman, Beaudoin & Wright, 1994).

Motive management processes themselves require management. *Self-monitoring mechanisms* can detect global states of the management processes, such as noticing that new motivators are surfacing faster than they can be processed (described as ‘busyness’ in (Beaudoin, 1994)). This could lead to a raised filter threshold. Other states worth detecting include repeatedly considering the same goals or problems without making progress in solving them (described as ‘maundering’ in (Beaudoin, 1994)).

It is acknowledged that this design sketch is speculative, vague and subject to revision in the light of implementation problems or empirical evidence. It is speculative in that empirical checking has not yet been attempted (and may be very difficult.) A small subset of the architecture has been implemented (Wright, 1994) and tested in a simulated domain, using very simple management processes. A more elaborate implementation is in progress, using a toolkit developed at Birmingham (Sloman & Poli, 1995). Practical problems encountered during implementation have fed and will feed back into the design stage. This is an iterative process, which is at an early stage. Nevertheless we think the architecture already has explanatory power. For example, it can begin to explicate certain characteristics of emotional states.

Perturbances

Both a grieving person and an intensely joyful person are partly ‘out of control’; that is, *if* they wanted to direct their thoughts to other matters, they would both find it extremely difficult to do so. Explaining this requires an analysis of what it is to be in control. The ordinary notion of 'self-control' is not a unitary concept admitting of a unique analysis. Like many mental concepts it covers a variety of cases and further study is required to investigate how many of them can be accommodated within the design-based framework.

The term *perturbant* is reserved for a state in which partial loss of control is due to the continual surfacing of postponed or rejected, or unwanted, motivators (Beaudoin, 1994), or possibly disruptive thoughts, images, and the like (e.g., a catchy tune that won't 'go away'). Such disruption can interfere with the management of other, important goals. This is the type of information processing state that the *afp* theory posits as characteristic of many emotional states (Simon, 1967; Sloman & Croucher, 1981; Sloman, 1987; Sloman, 1992). It is what the filter normally prevents.

Perturbant states differ in several dimensions: duration; whether the source is internal or external; semantic content (what is referred to); type of disruption (it could be due to a goal, thought, or recollection); effect on management processes; frequency of disruption; positive or negative evaluation (compare grieving with being unable to stop thinking about the victory one has recently won - grieving and gloating have much in common); how the state develops; whether and how it decays; how easily it can be controlled, and so on.

Perturbances (like 'thrashing' in an overloaded computer operating system) are side-effects of mechanisms whose major role is to do something else (just as thrashing arises from the paging and swapping mechanisms in the operating system). Perturbances arise from the interactions between (a) resource-limited attentive processing, (b) a subsystem that generates new candidates for such processing and (c) a heuristic filter mechanism. These design elements arise from the requirements for coping with complex and rapidly changing environments. Perturbances do not arise because of some special perturbance generating (or emotion generating) mechanism. Thus it is misguided to ask what the *function* of perturbant states is or to postulate a perturbance mechanism.

In (Wright, Sloman & Beaudoin, 1996) we used the concept of perturbance to provide a design-based analysis of loss or grief. However, we were unable to provide a satisfactory account of perturbant states that possess a pleasure or unpleasure dimension (compare grieving with glee). It was also unclear whether an information processing architecture as described above could account for types of mental pleasure and pain. In the next section this difficulty is described and a definition of valency provided.

VALENCED PERTURBANCES

It is taken as given that a full understanding of emotional states will include an understanding of internal, mental states that may not be predictably related to behavioural phenomena, facial expression, or general state of arousal. In this context, it is useful to analyse the phenomenology of mental states. Evidence from phenomenology examines the structure and function of mental states via introspection. An assumption is that introspection can provide information that is useful for theory construction, and that the phenomenological structure of mental states matches aspects of their functional implementation; in other words, it is possible that functional and phenomenological analyses can partly converge. However, it is admitted that introspection depends on self-monitoring capabilities that may be limited to very superficial aspects of internal processes, and like external perception may often be wrong. The long term test of such an approach will have to be in the context of a far more complete theory and new evidence. In the absence of detailed theories of the underlying mechanisms, care is required when employing phenomenological concepts.

In this section it is shown that there are aspects of mental states that are non-intentional, in the sense that they do not refer, and quantitative, in the sense that they vary in intensity.

Thinking and feeling

Cognitive representations can denote or refer to states-of-affairs. For example, an agent within a simulated domain may possess information about other agents in the environment, including their type, location, or speed. Such information control sub-states[2] of the animat are causally linked to their referents: the representation of the speed of agent A will alter if it is perceived that agent A has altered its speed. This is a simple example: referential links can be very indirect in more complex information processing systems. The principle, however, is conceptually clear and forms a basis of information processing theories of mind: *thinking refers*.

Emotions, desires, and pleasures and pains, differ from 'cold' cognition: they can be 'hot', often involving feelings of pleasure or unpleasure with associated intensities. Unlike 'straightforward' representational thinking, an emotional state sometimes has *both* a representational content, for example, a state of happiness *about* passing one's exams, and a hedonic, or *valenced* content, such as the particular form of intense pleasure one is experiencing. The hedonic component does not represent a state-of-affairs: *feelings just 'are'*. This is not to say that 'feelings' do not play a functional role, or are not causally linked to other processes.

A preliminary division can be made between *dispositional* and *occurrent* emotional states (Ryle, 1949; Green, 1992). A dispositional state is a latent state that may manifest in appropriate circumstances, such as the brittleness of a wine glass, whereas an occurrent state is a *running* state, such as the process of a wine glass breaking. For example, a man who has lost a parent may function normally at work (dispositional grief), only to break down in the evening (dispositional state manifests as an occurrent emotion). There are often two distinguishable components of an occurrent emotional state: *intentional* and *non-intentional*. The intentional component of an emotional state is what the state is about. A person is angry, disappointed, or ecstatic 'about' a perceived state of affairs. This state of affairs may exist in the agent's environment, or entirely within cognition (as in the case of the mathematician irritated with himself for being unable to solve an equation). The intentional component has representational content. Some philosophers wish to restrict intentionality to systems that are 'conscious'. No such restriction is intended here. It is

[2] A control sub-state is a general term for a subcomponent of an architecture bearing information (Sloman, 1996) that has relations of control to other subcomponents or parts of an environment. For example, the curvature of a thermostat's bi-metallic strip is a substate of the thermostat architecture that contains information about the ambient temperature of the room. The cumbersome 'control sub-state' is simplified to 'substate' in this paper.

hoped that a design-based approach will eventually unpack the philosophical concept of intentionality and provide a taxonomy of forms of representation that is grounded in an understanding of the variety of functional roles of substates in control systems.

The non-intentional component of an emotional state is often referred to as its 'hedonic tone', *feeling*, or *valency*. 'Feeling' is an ill-defined word, for it can cover such diverse sensations as one's cheeks burning with embarrassment, an itch on the left ear, or the mental happiness associated with triumph. Also, the word 'feeling' can be used in many other contexts, such as feeling like having a cup of coffee, or feeling in the mood to dance and so forth. 'Hedonic tone' is similarly semantically overloaded: it can be used to refer to the enjoyable sensation of a full stomach after a large and hearty meal. The word 'valency', if given a suitable definition, can avoid such confusion. Before giving such a definition we require some more distinctions.

A division can be made between *physiological* forms of pleasure and unpleasure, and *cognitive* valency. For example, the (self-monitored) 'itchiness' on my left ear is a form of unpleasure linked to information concerning bodily location. In contrast, the mental pain of intense grief is a form of unpleasure linked to information about a loved one's death. There can be no 'pain receptors' for this kind of unpleasure, unlike the nerves that detect a pin pricking one's finger. To illustrate: an athlete may be experiencing the occurrent emotional state of triumph while standing on the winner's podium. The intentional component of her state includes thoughts pertaining to the achievement of long-term goals and, for example, information about a rapid heart-rate and the warm sun beating on her brow; the non-intentional component *includes* the pleasurable sensation linked to the warm sun on her brow, *and* a valenced state of cognitive pleasure not located on or in the body but linked to her athletic success. The form of pleasure linked to information concerning the warm sun on her brow plays a control role: it serves to preserve the activity. This kind of causal role of 'happiness' (and 'sadness') control signals has also been identified by (Oatley & Johnson-Laird, 1985; Oatley, 1992). Their communicative theory is discussed in section 5.1.1.

Unlike the intentional component of an emotional state, valency does *not* represent a state of affairs. It can differ qualitatively in only a very restricted sense, that is it can be *either* pleasurable or unpleasurable, and allow quantitative degrees of *intensity*: the valency can be very unpleasurable or only mildly so. Very intense valenced states tend to 'grab hold' of attention and are difficult to ignore or suppress. From a phenomenological perspective, valency is a 'brute fact'[3] of one's present state, unlike beliefs that can be true or false, or goals that have been achieved or not – valency is neither true or false, nor does it represent a state of affairs to be achieved or avoided; rather, it simply 'is'. (This is not to imply that it is a scientifically inexplicable phenomenon; on the contrary, it is the aim of this paper to show that such phenomena can be explained in terms of the functional role of control signals in an information processing architecture.)

Intense forms of 'happiness' and 'sadness' provide the clearest examples of valency. The discussion, therefore, is currently restricted to these emotional states. Particular examples of happiness are triumph, glee, joy, ecstasy, gladness, (occurrent) love; and of sadness, despair, disappointment, grief, and sorrow. For the sake of brevity, it will be stated that a person is 'happy' when a desired goal is achieved, and 'sad' when failure occurs in achieving a desired goal. This is an oversimplification: concrete emotional states are rarely this straightforward or as simple. Also, not all goal success or failure leads to valenced states: intuitively, the goal must be important to generate valency. Particularly important goals are those that occur during the formation and destruction of personal attachments. Section 5.1.2 attempts to explain the difference between attachment goals and other goals by comparing the emotional states of grief and disappointment. Valency can be preliminarily defined as follows:

> *Valency* is a form of cognitive pleasure or unpleasure not linked to information concerning bodily locations, and is a quantitatively varying, non-intentional component of occurrent emotional states of happiness or sadness. Valenced states are contingent on the success or failure of subjectively important goals; for example, processes of attachment and loss often involve valenced states.

[3] This term borrowed from (Chalmers, 1996).

This definition of valency is a special case of a more general notion. For example, valency as defined here is not to be confused with other forms of pleasure or unpleasure that are shorter-term control states involved in whether current activities are preserved or terminated; neither should it be confused with 'values' in general, such as more interesting, qualitative, or symbolic, affective dispositions towards states of affairs. Instead, valency is 'achievement pleasure' or 'failure unpleasure' that occurs when certain highly important concerns are met or violated, such as the concerns that develop during the formation of personal attachments. A causal role for valency is provided in section 5.1.

Summary: the need to explain valenced perturbances

The architecture described above can generate perturbant states. Many emotional states, such as intense grief and triumph, involve a partial loss of control of attention. However, they differ in two important respects: first, grief is contingent on the violation of concerns and may generate second-order motivators to remedy the situation, or suppress thoughts, or stop the 'pain' and so forth, whereas triumph is contingent on the satisfaction of concerns, and is therefore less likely to immediately generate new motivators; second, both triumph and grief involve a valenced component, which is a form of mental pleasure or unpleasure: grief can be very painful, whereas triumph can be very pleasurable.

The *afp* theory needs to account for the valency of perturbant states. The insistence of a motivator defines its ability to surface and obtain attentive processing resources; however, insistence is not linked to valency, nor is it clear why it should be as, for example, many 'attention grabbers' are neither pleasurable or unpleasurable (consider a loud bang). Intuitively, however, the intensity of the valenced component of some emotional states can be correlated with the allocation of attention. The intense mental pain of grief is very hard to ignore or control.

Many information processing theories of emotion tend to avoid an explanation of the valenced components of emotional states by concentrating on the semantics of representational components (e.g., Dyer's

BORIS system (Dyer, 1987), Frijda and Swagerman's ACRES system (Frijda & Swagerman, 1987), and Pfeifer's FEELER system, reviewed in (Pfeifer, 1994)). Alternatively, feelings are brushed under the physiological carpet by assuming that all valenced states arise from perceptions of bodily states. For example, the following quotations from Herbert Simon's seminal paper on motivation and emotion (Simon, 1967), outlines a view of 'feelings' that closely resembles William James' peripheric theory of the emotions: '... sudden intense stimuli often produce large effects on the autonomic nervous system, commonly of an "arousal" and "energy marshaling" nature. It is to these effects that the label "emotion" is generally attached'; and '... the feelings reported are produced, in turn, by internal stimuli resulting from the arousal of the autonomic system'. Although Simon's functionalism clearly understands the need for control in addition to representation, it is difficult to conceive how this view of valency could account for the mental pain associated with, for example, grief, which does not necessarily require bodily arousal or disturbance.

Therefore, there appear to be at least two reasons why explanations of valency are generally absent from cognitive theories: first, valenced components appear not to conform to the representational model that supports cognition; and second, their possible functional role is unclear: what can these states possibly *do* if they do not represent? Why do such diverse and complex states such as happiness, sadness, glee, triumph, grief, despair, intense disappointment and so forth have valenced components?

Introspection suggests the existence of forms of cognitive pleasure and unpleasure that are non-intentional and vary in quantitative intensity. In the next section a requirement for *adaptivity* is considered. This is a preliminary step before a theoretical explanation of valency can be given in section 5.1.

DESIGNS WITH A QUANTITATIVE UNIVERSAL REPRESENTATION OF VALUE

In this section arguments for the existence of a quantitative universal representation of value in motivational subsystems are examined, including evidence from design, considerations from reinforcement learning, artificial intelligence systems, and economic systems. Such an investigation is in the spirit of (Simon, 1981): that is, an endeavour to try and discover generic design solutions within complex systems so that the solutions from one type of system may be usefully transferred to the design or analysis of another. The evidence is then synthesised, generating an 'economic' design hypothesis.

Designs with internal forms of value

Considerations from design examine the abstract, information processing properties of systems designed to meet various requirements. It is shown that designs that operate in niches requiring adaptive behaviour utilise internal forms of value. Value is internally relational and is to be contrasted with belief-like and desire-like substates that refer to aspects of the niche or internal states. Reinforcement learning algorithms employ quantitative forms of internal value. Economics systems show the potential usefulness of quantitative representations of value in highly complex systems with many levels of control.

The emergence of value in design-space

The development of negative feedback control systems demonstrated that simple, materially embodied systems can have substates with different functional roles, in particular 'belief-like' and 'desire-like' substates (McCarthy, 1979; Sloman, 1993b; Dennett, 1996; Powers, 1988; Braitenburg, 1984). For example, the belief-like substate of a thermostat is the curvature of its bi-metallic strip, which alters in accordance with the ambient temperature of a room; the desire-like substate is the setting of the control knob. Negative feedback ensures that the temperature of the room stabilises around the control knob setting: the thermostat 'acts' in the world to achieve its 'desire'. Of course, a thermostat does not have sufficient architectural complexity required to support human beliefs and desires, but it is an illustrative 'limiting case' (Sloman, 1993a) of a system that controls its environment with reference to an internal norm.

A new feature and a new requirement can be imposed on the simple thermostat. The new feature is the effect of the thermostat's 'actions' are now *uncertain*; that is, from the perspective of the thermostate designer the output signal has (initially) unknown effects on the temperature of the room. The new requirement is that the thermostat adapt and improve its behaviour over time: it is required to learn the most efficacious output signals for controlling temperature. A change in requirements is a movement in niche-space. The new type of thermostat will need to try out actions, monitor the effects of those actions, evaluate the results, and then select and retain those actions that work best. This is the corresponding movement in design-space.

In the abstract, the new niche requirements can be satisfied by a *selective system* (Pepper, 1958). A selective system has three components: (i) a *trial generator*, which is any mechanism that generates a variety of functions to produce outputs for particular inputs, (ii) an *evaluator*, which is a mechanism that evaluates the results of using particular functions to generate trials, where evaluation occurs through comparison to a norm, such as success in bringing the temperature closer to the desired setting, and (iii) a process of *selection*, which retains those functions associated with 'good' evaluations for future use, while discarding others. Selective systems implement the well-known generate, test, and select cycle. Specific examples of selective system improve their behaviour over time (cf. Darwinian evolution and genetic algorithms).

Many possible designs could meet the specification of a selective system, yet all such designs evaluate their trials. In the adaptive thermostat example, if $trial_1$ brings the room temperature closer to the norm, whereas $trial_2$ does not, then the substate that produced $trial_1$ has greater *value* to the system and will be retained for future use. In whatever form internal value is implemented within the control system it will function to order trials by specifying a relation between them, for example $trial_1$ *is_better_than* $trial_2$.

Just as there is a great variety of belief-like and desire-like substates in design-space, there will be a great variety of value-like representations in control systems. It is a highly abstract concept, and is as important as the information processing definitions of 'beliefs' and 'desires' (when these concepts are understood to have been abstracted from folk psychology and applied to control systems in general). Unlike beliefs, which have a 'world-to-mind' direction of causality, and desires, which have a 'mind-to-world' direction of causality, value is internally relational, having 'mind-to-mind' causal interactions, and refers, in an impoverished sense, to the 'goodness' or 'badness' of substates. Value appears in design-space when a requirement for adaptivity is imposed. It orders internal components with reference to normative criteria within a selective system. In the next section a particular form of value is examined.

Reinforcement learning algorithms

The study of reinforcement learning (RL) algorithms is an active area of research (Kaebling, Littman & Moore, 1995). RL algorithms are selective systems as defined above. RL is a type of trial and error learning, and holds out the promise of a way of programming control programs for agents by reward and punishment without the need to specify *how* a task is to be achieved. The main design problem to be solved in reinforcement learning is the *credit assignment* problem, which is the problem of 'properly assigning credit or blame for overall outcomes to each of the learning system's internal decisions that contributed to those outcomes' (R. S. Sutton, quoted in (Cichosz, 1994)). More precisely, RL involves learning functions defined on the state and action space of a task, driven by a real-valued reinforcement signal. The details of how this is achieved depends on the particular function representation used. In the abstract, RL involves constructing a function, called the policy function, $f : X \longrightarrow A$, where X is the set of possible learner-environment states and A is the set of possible actions the learner can perform, such that the cumulative reinforcement (r) over time (t), $\sum_{t=1}^{\infty} r_t$, is maximised. Reinforcement is derived from conditions that define the utility of particular states, and can be thought of as the abstract requirements of the task. In many implementations f is implemented as a set of candidate functions (e.g., $f(x_i) \longrightarrow a_j$, where $x_i \in X$ and $a_j \in A$), each of which have an associated value that determines the likelihood of it being selected for use and determining behaviour. The form of value in RL algorithms is a scalar quantity (i.e., is not decomposable).

Holland's classifier system (Holland, 1995; Holland, 1975; Holland, Holyoak, Nisbett & Thagard, 1986; Riolo, 1988) is a concrete example of a RL algorithm. It consists of a *performance system*, and *credit-assignment* and *rule discovery* algorithms.[4] Rule discovery will be ignored in this summary. The performance system consists of a *classifier list* that consists of a set of condition-action rules called *classifiers*, a *message list* that holds current messages (as in the working memory of production systems), an *input interface* that provides the classifier system with information about its environment in the required form, and an *output interface* that translates action messages into world events. The *basic cycle* of the performance system matches messages in the message list (including sensory messages) with classifiers, which then post their actions 'back' to the message list. Many classifiers may become active and fire in parallel. Any current action messages are sent to the output interface.

Reinforcement learning is achieved via a *bucket-brigade* (*bb*) algorithm. This algorithm introduces competition between classifiers based on a quantitative 'strength'. Each classifier that has its condition activated by a message *bids* to post its action part to the message list. Only the top subset of highest bidders on each cycle are allowed to post their actions. The bid of a classifier depends on its strength, which is a measure of the classifier's 'usefulness' to the system. The higher the strength of a classifier the more likely it will win the competitive bidding round and post a message.

The strength of classifiers specifies a probabilistic partial ordering on the classifier list. The ordering is partial because only some classifiers will bid for the same message. The ordering is probabilistic because the classifier selection mechanism is stochastic. For example, both classifiers c_i and c_j match message m; c_i has strength s_i, and c_j has strength s_j, with $\mathrm{s}_i > \mathrm{s}_j$. In competitive bidding for message m, c_i will out bid c_j more often than not: there is a probabilistic total ordering on the set $C = \{c_i, c_j\}$ (an example of the *is_better_than* relation). Other classifiers in the classifier list, such as c_k, *never* compete against any $c \in C$; consequently, no ordering holds between them. Strength is internal economy alone. It does not represent anything within or external to the classifier system; rather, it specifies a *internal*

[4] This specification abridged from (Riolo, 1988). Readers unfamiliar with classifier systems should consult (Holland, 1995).

relation between classifiers. It is this property of strength that helps to make the classifier system a *domain-independent* learning algorithm: the representation of utility does not alter from domain to domain.

The behaviour of the classifier system, therefore, can be modified by changing the strengths associated with individual classifiers. If the strength of the classifiers that tend to lead to 'useful' behaviour can be increased, and the strength of the classifiers that tend to lead to 'useless' behaviour can be decreased, the system will learn to produce more useful behaviour. The *bb* is designed to bring about these types of changes in strength.

The basis for the *bb* is information from *reinforcement mechanisms* about whether the classifier system as a whole is behaving correctly (as defined by the designer, or, if complete classifier systems are evolved in a simulated domain, by the niche fitness function). This information derives from *rewards*; that is, the system will receive positive reward when it behaves correctly and negative reward when it behaves incorrectly. When a reward is received the *bb* adds the reward value to the strength of all classifiers currently active, thereby changing the strength of classifiers *directly* associated in time with useful behaviour. Also, when a classifier is activated it 'pays' the amount it bid to the antecedent classifier that produced the message it matched. The strength of the active classifier is decreased by its bid amount. In this way, the *bb* acts to increase the strength of classifiers *indirectly* involved in the production of useful behaviour (e.g., a recently rewarded classifier will pay the same proportion but a higher amount of its strength to antecedent classifiers in subsequent cycles). The *bb* allows reward to 'circulate back' through the system. Chains of high strength classifiers, performing useful computation, can emerge from such a scheme.

The scalar strength associated with each individual classifier is a form of value. The *bb* was originally inspired by an economic metaphor (Holland, Holyoak, Nisbett & Thagard, 1986), in which classifiers are agents (consuming and producing messages) who possess a certain amount of money ('strength' or value) which they exchange for commodities at the market (the message list or blackboard). Much as money mirrors the flow of commodities in a simple commodity economy, *value mirrors the flow of messages* in the learning classifier system. Both money and value transmit feedback from the ultimate consumers, be they reinforcers or the needs of individuals. Whether the implementation of the classifier system is truly parallel (with perhaps separate processors for each classifier), or only simulates parallelism, *the value of a classifier is an ability to buy processing power*. A high value classifier will be more likely to win bidding rounds, be processed, and post its action part. For example, an internal sensory message may match the first in a chain of high value classifiers that instigate an action sequence. The high value of such a processing chain will make it unlikely that other rules will out bid and switch processing to other ends.

Classifier systems have been extensively used in the ALife community (Steels, 1994), for example in developing control programs for robots through supervised learning (Dorigo & Colombetti, 1993) and discovering paths through mazes (Donnart & Meyer, 1994). However, a classifier system is limited in many ways. It does not have an explicit memory store. It tends to be an entirely reactive system with no representation of goals. It does not anticipate, or perform prior search within a world model before acting. In real-world applications it can be difficult for a classifier system to learn appropriate behaviours (Wilson & Goldberg, 1989). Also, the classifier system is not a fixed architecture but continues to evolve, for example the recent introduction of rule discovery based on the accuracy of classifier predictions (Wilson, 1995). Despite this, the classifier system is an existence proof of a selective system that uses a quantitative form of value to meet a requirement for adaptivity.

Economics and exchange-value

As stated, the design of a classifier system was inspired by an economic metaphor. Social systems are often used as a model for cognitive systems, for example (Minsky, 1987), because they share design features; that is: (i) a set of mutually connected, interacting subcomponents that are able to perform work (e.g., neurons that process or people that labour) (ii) that can function as both producers and consumers (e.g., the input and output of information, or the consumption and production of commodities), (iii) operating within a social division of labour (i.e., a functional specialisation of subcomponents) (iv) that exhibits fine-grained parallelism (i.e. the subcomponents function relatively autonomously and concurrently), (v) which needs to be coordinated by mechanisms for the exchange and distribution of products (such as the propagation of information, or free market mechanisms for commodity exchange in current economic organisation). In addition, economic systems (vi) allocate scarce resources, be they limited labour resources or commodities in restricted supply. It is likely that a society of mind will require an economy of mind, and various economic methods and concepts will have new application.

In the abstract, economic systems are selective systems: the trials are the various concrete labours that produce commodities, the evaluatory mechanisms are the various needs and demands of individuals, and selection occurs through the buying and selling of commodities. The value in economic systems is exchange-value (Marx, 1970), which is associated with commodities, and can be stored and exchanged in various forms, such as gold, paper money or virtual currency flows.

In *The Society of Mind* (1987) Marvin Minsky discusses the possible existence and functional role of 'currency' in cognitive systems.

> ... a group of agencies inside the brain could exploit some "amount" to keep account of their transactions with one another. Indeed agencies need such techniques even more than people do, because they are less able to appreciate each other's concerns. But if agents had to "pay their way", what might they use for currency? One family of agents might evolve ways to exploit their common access to some chemical that is available in limited quantities; another family of agents might contrive to use a quantity that doesn't actually exist at all, but whose amount is simply "computed". I suspect that what we call the pleasure of success may be, in effect, the currency of some such scheme.
>
> M. Minsky, *The Society of Mind*, p. 284 of (Minsky, 1987).

Minsky also points out that comparisons based on quantitative differences will be limited because they ignore qualitative differences, such as the loss of information that occurs when qualitative reasons for success or failure are reduced to directly commensurable numbers. This is an important point and has implications for the evolution of higher and more advanced evaluation mechanisms.

However, very complicated systems, with many levels of control, such as human economic systems, have developed currency flow to regulate some of their transactions. A money-commodity has a number of useful properties in this context. It is *universal*, in the sense that it is recognised by all economic agents. It is *functionally determined* as it admits of no local interpretation but is compared with other quantities of value. In other words, money is globally semantically determined in a society of numerate agents. Related to this is that currency exchanges only require relatively *simple operators*, such as addition, subtraction and numerical comparison. No sophisticated local machinery is required to mediate transactions. It has *multiple uses*. Its exchange incurs relatively *low communication costs* between agents, when compared to bartering systems, for example. It can propagate constraints over time and distance affording *distal connectivity* and the selection of *legal exchanges* that reinforce social cooperation: for example, a particular agent will not be able to acquire a commodity without first expending labour that has sufficient value to other agents (in abstract, simple commodity economies at least). In this way, the flow of currency implements a local and distributed solution to a global coordination problem.

If the form of value is an indestructible substance, such as gold, it can be *accumulated* and stored indefinitely. It is a domain-independent, or, better, *internally relational*, representation because it is compared to other quantities of value: changes in the kinds of labours performed and commodities produced within economic systems do not alter the functional role of money. Finally, the ubiquitous use of money in economic systems introduces global *supply and demand* dynamics that allow the system to adapt to changes in productivity and consumption. Information-theoretic analogues of some of these properties may be useful for coordinating adaptive, largely parallel information processing systems. For example, some recent work in AI uses economic ideas for resource allocation problems (Wellman, 1995), including allocation of processing time, and reasoning about plans (Doyle, 1994).

In summary, human economic systems constitute an existence proof that complex processing systems can be regulated by exchanges of a quantitative representation of value.

Summary of the evidence from design

An *is_better_than* relation appears in design-space when types of selective systems attempt to meet a requirement for adaptivity. For example, a behaviour producing substate that is selected over and above other substates has greater value or utility to the system. Belief-like and desire-like substates are relations between niche and system, and are to be contrasted with forms of value that constitute internal relations between parts of the system. There are many possible forms of value. Some may be implicit and not explicitly represented.

Existing designs for reinforcement learning algorithms employ an explicit quantitative representation of value (credit or strength) that is stored with behaviour-producing substates. A concrete example of such a system is Holland's classifier system, in which value plays a role as an ability to buy processing resources and dispositionally determine the behaviour of the system. In addition, economics systems also use a quantitative representation of value, which suggests that such a design feature may be useful for systems composed of many interacting parts operating in parallel that need to be adaptively coordinated.

A design hypothesis

By ignoring implementation details it is possible to extract design principles from the selective systems considered. These principles may be of use in understanding other systems that share design features. Reinforcement learning algorithms, such as Holland's classifier system, and human economic systems use a quantitative form of value that is exchanged between substates to perform credit assignment. This design feature is the presence of a *universal, quantitative representation of value.* It is universal in the sense that all the substates of the system are conversant with its use, and quantitative in the sense that it is a non-decomposable, scalar signal. This form of value *circulates* within these systems. A particular example of circulation of value is the bucket-brigade algorithm, where a substate is a single classifier, the system is the set of competing and cooperating classifiers, and substate products are messages with semantic content. The *bb* adaptively alters the ability of classifiers to *buy processing power* and determine the behaviour of the system. Currency flow in economic systems performs a similar function. For example, a producer who makes a profit will have more money to employ more people (to buy processing power directly) and more raw materials (to buy the results of prior processing). Individual profits and losses regulate this ability to commandeer and allocate social resources.

Interest devolves on explaining and understanding valency and valenced perturbant states that can occur in human brains and be simulated in suitably designed agent architectures. To explain these states the following design hypothesis is proposed. It is an 'economic' hypothesis because it is a simple design feature based on an analysis of systems that adaptively allocate resources.

> **An economic design hypothesis:** A motivational subsystem makes use of a circulation of value mechanism to perform credit-assignment. Circulation of value involves (i) altering the dispositional ability of substates of the system to gain access to limited processing resources, via (ii) exchanges of a quantitative domain-independent representation of value that mirrors the flow of substate products. The possession of value by a substate is an ability to buy processing power.

The design hypothesis leads to a set of high-level requirements that are fully described in the next section. The motivational subsystem considered is one that is primarily concerned with developing social attachments[5] to

[5] The term 'attachment' is generalised from Bowlby's theory of attachment (Bowlby, 1979; Bowlby, 1988), which is primarily concerned with the behaviour of infants towards caregivers. 'Attachment' in this document, however, applies to the development of a social relationship, which, if it ended in circumstances that contravene the desires of the subject, would cause emotional upset, including protest, despair and eventual detachment.

caregivers, parents, sexual partners and so forth. The subsystem learns means of satisfaction (goals, plans, threat and opportunity detectors, generactivators etc.) from basic, untaught conditions of satisfaction (untaught, 'attachment' reinforcers). The relatively simple design principle of a motivational 'common currency' can account for valency and improve upon existing information processing theories of (some) emotional states.

The design hypothesis is a hypothesis because it is a statement about the world that can be falsified. It states something about the *functional* organisation of a motivational subsystem at the level of information processing. It states nothing about the implementation of such functionality: it could be realised in rule-based systems, neural networks, multi-agent systems and so forth. The hypothesis would be strongly falsified if future neurological or computational studies of human brains reveals that their functional organisation does not include a quantitative representation of value with properties as defined. However, this endeavour would be similar to examining the switching of logic gates in a microchip in order to discover what algorithm is running on a personal computer. A more fruitful route to falsification is an exploration of design-space, in particular the development of agent architectures that meet a requirement for adaptivity through the use of forms of reinforcement learning. Such an exploration could reveal that circulation of value mechanisms cannot meet the requirements, or that they are necessary but not sufficient, or they are one of a range of possible alternatives, or they work only for certain types of architecture and so on. Ultimately, it is an empirical question. However, the hypothesis is sufficiently general that it applies to a wide class of systems. Therefore, in the next section the hypothesis is given more content by providing related design requirements and situating it in an agent architecture.

PRELIMINARY DESIGN REQUIREMENTS FOR A COMPUTATIONAL LIBIDINAL ECONOMY

Requirements analysis is the first stage of software engineering methodology. Requirements define what the intended system is supposed to do but not the details of how it will be done. The latter is the problem of design and implementation. Requirements pose a design problem without providing a solution: there are many possible designs that could meet the requirements.

The requirements listed below are *design requirements*, as they stipulate the high level features of possible designs for a computational libidinal economy. It is not yet specified in sufficient detail to be called a 'design'. For example, the simulation of a computational libidinal economy would require advances in reinforcement learning architectures that include more complex motive management processes. The design requirements constitute an abstract theory for a computational libidinal economy (*cle*) that instigates and controls the formation of attachments (particularly sexual) between humans. It is integrated with the agent architecture that was described in section 2.2.

The design requirements fail to stipulate the kinds of substates that constitute the libidinal economy: details of the mechanisms that produce information with semantic content, what that information represents, how it is transformed, and what relations these mechanisms have with themselves and other cognitive systems are missing. The requirements describe a *cle* composed of black boxes, or economic agents without content. This is a deficiency from the perspective of building a working system, but is less of a problem from the perspective of explaining valency, valenced perturbances, allocation of attention and relations to reinforcement learning. A fuller theory would be sufficiently specified as to be implementable. The problem of specifying the information processing 'agents' within the libidinal system will depend on advances in artificial intelligence systems, in particular adaptive multi-agent systems.

Finally, the design requirements oversimplify drastically, and hide much unknown complexity. This is currently unavoidable when first attempting to understand a highly complex information processing system, especially when the computational processes are unobservable.

The design requirements for a *cle* are as follows:

1. *A libidinal selective system.* The libidinal selective system is a cognitive subsystem. Its main purpose is to develop social attachments to others, such as caregivers or partners.

(a) *Means of satisfaction.* To develop an attachment the libidinal system is required to construct a superstructure of motivational substates that constitute the means of satisfaction of various fundamental attachment goals. The motivational substates can generate motivators to higher level systems, such as attentive processing. The libidinal system is an automatic, largely preattentive process. composed of a 'society' of competing and cooperating substates. All substates have conditions of activation, which are patterns that can match semantic products (much like the condition part of a production rule). Substates, when activated, output semantic products based on internal state and their current input. For example, some substates may have conditions of activation that match sets of beliefs (which, in the current agent design are assumed to be stored in a globally accessible world model (Beaudoin, 1994; Sloman, Beaudoin & Wright, 1994)).

(b) *Untaught conditions of satisfaction.* Untaught reinforcement mechanisms define the fundamental attachment goals by specifying various conditions of satisfaction that have been selected by evolution. Example conditions of satisfaction are orgasm, proximity of mate, positive emotional signals from opposite sex (e.g., facial expression, laughter, interest etc.) and so forth. It needs to be stressed that attachment in humans involves much more than reinforcement learning. For example, it appears to include imprinting-like mechanisms and the construction of predictive models of the attachment figure. However, the libidinal selective system is specifically concerned with reinforcement learning. Its relations to other kinds of learning is an open question.

(c) *Learnt conditions of satisfaction.* A subset of the means of satisfaction are learnt conditions of satisfaction, or taught reinforcers. These substates reinforce subgoals, which are useful 'landmarks' towards the eventual achievement of untaught conditions of satisfaction. Learnt reinforcers inherit their reinforcing properties from untaught reinforcers.

(d) *A selective cycle.* The libidinal system is a selective system and is therefore required to perform three functions: generate substates that are candidate means of satisfaction, evaluate those substates with reference to internal norms, such as utility in satisfying untaught and learnt conditions of satisfaction, and select better substates for future use, while deselecting others. This is achieved by reinforcement and processes of selection upon value. For example, a substate involved in the production of behaviour that satisfies a reinforcer will gain in value and be more likely to dispositionally determine the behaviour of the system in the future. However, actual reinforcement will not be the only form of evaluation in human-like architectures: depending on domain knowledge, evaluation could occur with reference to a world model sufficiently detailed to predict the consequences of actions. For example, a libidinal substate may generate a motivator to a higher level system that is able to predict the likely consequences of adopting the motivator as an intention. A motivator with high expected reward may be more likely to be adopted by management processes than a motivator with low expected reward. These kinds of relations between reinforcement and more complex forms of motive management will only be resolved through the further exploration of agent architectures.

(e) *Libidinal generactivators.* The complexity of the substates of the system will vary. There will be hierarchical relations of control between substates. Some substates are libidinal generactivators, whose semantic products are candidate motivators for attentive processing. These motives may attempt to meet conditions of satisfaction of reinforcers, or may detect threats or opportunities relevant to current attachment plans. They are examples of Frijda's *concerns* (Frijda, 1986). For example, particular libidinal generactivators may detect threats to attachment plans, such as interest from other sexually active males or females towards the loved one.

2. *A conative universal equivalent.* The universal, quantitative representation of value within the libidinal selective system is the conative (= 'motivational') universal equivalent (*cue*). It functions as a universal means of exchange between the substates of the system. It can be exchanged by a substate for the semantic products of a producer substate, like money is exchanged for commodities in an economic system. *Cue* is stored with each individual substate. It is assumed that the initial allocations of *cue* to untaught substates are genetically specified. See 4a for the sources of *cue*.

(Note: *cue* is a universal means of exchange *within* the libidinal system. It is an open question whether circulation of value can be generalised to other motivational systems. For example, homeostatic and simpler feedback motivators, such as maintaining temperature, removing waste, altering body posture and so forth, do not involve the kind of valenced perturbant states characteristic of processes of attachment (see section 5.1.2).)

3. *Possession of* cue *is an ability to buy processing power.* The possession of *cue* by a substate is a dispositional ability to buy processing power. This can take a number of forms.

   (a) *Possession of* cue *is a dispositional ability to enter preattentive circulation. Cue* can be used to enter circulation within preattentive, automatic processing. A substate can exchange *cue* for the semantic product of another substate. If many agents compete for limited processing resources in the same context then a conflict resolution mechanism is required. The requirements simply state that the tendency to win all kinds of computational resources (e.g., processing time, the semantic products of prior processing, the ability to direct the processing of other substates etc.) is correlated with the possession of *cue*. There are many possible conflict resolution mechanisms – a simple example is provided by classifier systems.

   (b) *Possession of* cue *is a dispositional ability to construct motivators for management processing.* Libidinal generactivators exchange *cue* for semantic products that satisfy their conditions of activation. Consequently, if there is competition for information, generactivators with high *cue* will tend to win the competition and construct motivators. These become candidates for surfacing.

The production of a motivator by a libidinal generactivator *costs*. That is, there are 'prices of production' associated with the construction of a motivator and an attempt to enter management.

(c) *Possession of* cue *is a dispositional ability to construct motivators that surface and grab management resources.* This can be understood by considering the relations between *cue*, importance and insistence.

Insistence is a locally computed heuristic measure of the urgency and *importance* of a motivator (see section 2.2 and (Beaudoin & Sloman, 1993)). Motivators that lead to highly rewarding consequences are important. For example, libidinal generactivators high in *cue*, that is those generactivators that led to rewarding consequences in the past, will compute relatively high insistence values for the motivators they produce. The importance component of insistence is proportional to an expected reward extrapolated from past results. In other words, libidinal generactivators high in *cue* will have high dispositional powers to produce motivators that surface and determine attentive processing.

However, once a motivator has surfaced, management processes decide whether to adopt (but not schedule) a motivator based largely on more sophisticated calculations of the importance of a motivator. Major differences between (i) the local, preattentively computed importance component of a motivator's insistence, and (ii) attentively computed importance measures, will lead to perturbant states. For example, a libidinal generactivator may continue to produce insistent motivators even when management processes decide they are unsatisfiable, such as when the object of the motivator has died (see the account of loss in section 5.1.2).

4. *The exchange of* cue *mirrors the flow of semantic products within the libidinal selective system.* For a substate to enter circulation it must pay the substate that produced the semantic product it matches. Consequently, a single, local exchange of semantic information also involves a local exchange of *cue.* The antecedent substate, because it produced 'useful' information that was 'bought', receives a local 'reward', that is it gains an amount of *cue* from the buyer. The buyer, therefore, partially selects the producer. If the antecedent substate receives more *cue* than it paid to *its* producer it will gain in social power: the chain of production is profitable.

(a) Cue *derives from reinforcers.* The ultimate sources of value are the normative criteria of the libidinal selective system, that is untaught reinforcers. Secondary sources of value are learnt reinforcers. When the conditions of satisfaction of a reinforcer are met the substates involved in producing those conditions receive *cue.* Detailing this process further would require solving the temporal credit assignment problem.

(b) *Gain of* cue*.* A substate gains value by entering circulation and receiving more *cue* for the information it produced than it paid out to an antecedent substate for its preconditions. This case subsumes the situation in which the preconditions for a reinforcer (learnt or untaught) are met and the antecedent substates involved in the production of the preconditions are rewarded accordingly.

(c) *Loss of* cue*.* A substate loses value by entering circulation and receiving less *cue* for the information it produced than it paid out to an antecedent substate for its preconditions. This case subsumes the case where preconditions for a negative reinforcer (learnt or untaught) are met and the antecedent substates involved in the production of the (aversive) preconditions are negatively rewarded accordingly. There need be no 'sink' for lost amounts of *cue*: it is assumed that destructive computational operations can be applied.

(d) *Accumulation is 'reinforcement'.* The accumulation of *cue* by a substate is 'reinforcement' learning, in the sense that the substate will have increased dispositional ability to determine (internal or external) behaviour in similar contexts in the future.

(e) *Loss is deselection.* The loss of *cue* by a substate is the partial deselection of that substate. It will have less dispositional ability to determine (internal or external) behaviour in similar contexts in the future.

The exchange of *cue* is an important part of the reinforcement learning that occurs within the libidinal selective system. Substates that are adapted, in the sense that they are involved in the achievement of conditions of satisfaction, gain in value; those that are not, lose value.

5. Cue *is internal economy and has control semantics. Cue* is a form of value within a selective system. It is internally relational specifying an ordering of utility over substates. It has the control function of an ability to buy processing power. Unlike beliefs or desires it does not refer to any thing within the system, nor does it refer to anything external to the system. It is domain-independent.

The first corollary to these requirements is an explanation of valency, a certain form of mental pleasure and unpleasure.

Corollary: a circulation of value theory of simple affect

A computational libidinal economy that meets the requirements outlined above has two distinguishable components: an intentional and non-intentional component.

The intentional component of the *cle* is the set of substate products, in particular the motivators produced by libidinal generactivators. In contrast to *cue*, substate products have representational content: they are 'about' other things, be they states-of-affairs in the environment or within cognition. A concrete, if simple, example can be provided by the classifier system (for a fuller account see (Wright, 1996)). The substate products in this system are messages. Imagine an artificial frog embedded in an environment of real or simulated flies. The control program for *simfrog* is a classifier system with an adapted set of classifiers. The *simfrog* has an eye sensor, which forms part of the classifier system's input interface. The eye can detect a number of attributes of any fly within range. Attributes could include whether the fly is moving, what colour it is, its size and proximity. If a fly is detected the eye sensor posts a message to the message list that encodes this information. This sensory message is the result of a *mapping* between a state of the environment and a sub-state of *simfrog*, that is the message has representational content. Internal messages, less directly linked to sensing or acting, will have more complex representational roles within the system. The *semantics* of messages depends on the dynamic relationship between message and environment. For example, the sensory message may match a classifier that posts an action message that results in *simfrog* throwing its sticky tongue in the direction of the detected fly. The meaning of the message, therefore, would be an impoverished version of the imperative, 'eat that fly!'.

The non-intentional component of the *cle* is the circulation of value. The circulation of value is a pattern of flow of control, as opposed to semantic, signals [6]. Such signals have no semantic content and propagate around the system altering control flow.

Consider that self-monitoring mechanisms (section 2.2 and (Wright, Sloman & Beaudoin, 1996)) are required to monitor the circulation of the *cue* (value) occurring within the libidinal economy. The reasons why this may be necessary are not discussed here. The circulation of value from one moment to the next will involve a *net* exchange of value, which can be written as V_t, from matching substates to antecedent substates. The self-monitoring mechanism records each V_t over a specified time period, say $t = 1 \ldots n$, and displays the *change* in value, denoted δV_t, which is exchanged from one time step to the next, where $\delta V_t = V_{t+1} - V_t$. Again, a concrete example is provided by the classifier system. For example, *simfrog* may be in the process of learning how to catch and eat a fly. During this

[6] The following analogy may help capture the distinction. Imagine trains travelling on a complex network of tracks. Postal trains contain mail (semantic content) with destination addresses on the envelopes. These trains travel to the destinations and deposit the mail (the information). However, a different kind of train, a 'control signal' train, can travel through the network altering the points of the tracks. This has the effect of changing the topology of the network: trains will continue to deposit their mail but will use different routes.

process δV_t can be either:

- Positive, implying (a) a net increase in the utility of antecedent classifiers (substates), and (b) currently active classifiers are likely to lead to positively rewarding consequences;
- Negative, implying (a) a net decrease in the utility of antecedent classifiers, and (b) currently active classifiers are likely to lead to negatively rewarding consequences; or
- Zero, implying no net change in the utility of antecedent substates.

Therefore, the self-monitoring of circulation of value will display a rate of change of value with both sign and magnitude. Consider connecting the output of self-monitoring to *simfrog*'s skin, which can change colour. If δV_t is zero *simfrog* remains *green*, if δV_t is positive he displays *yellow* with an intensity $|\delta V_t|$, and if δV_t is negative he displays *blue* with intensity $|\delta V_t|$. When *simfrog* catches and eats a fly he will blush bright yellow as innate reinforcement mechanisms strongly positively reward antecedent classifiers. If *simfrog* possessed more sophisticated reflective capabilities he might wonder why he has beliefs that refer *and* an odd quantitative intensity that is either positive or negative but doesn't seem to be 'about' anything or serve any readily identifiable purpose. Depending on philosophical prejudice, one might be tempted to say that *simfrog feels* happy, sad or indifferent depending on circumstance.

Obviously, the example is a major simplification. But it shows that the monitoring of circulation of value in the *cle* can generate non-intentional control states, which can be either positive or negative, and vary in intensity. In addition, when a goal is achieved, such as the achievement of the conditions of satisfaction of a reinforcer, there will be a monitored increase in value. Similarly, if a goal of avoiding the conditions of satisfaction of a negative reinforcer fails, there will be a monitored decrease in value. Achievement or failure of certain fundamental goals as defined by reinforcers is linked to positive and negative exchanges of value respectively.

Therefore, another requirement can be added to the libidinal economy:

6. *Valency is the monitoring of a process of credit assignment.* The monitoring of circulation of value is the architectural process that gives rise to valenced states, that is forms of mental pleasure and unpleasure.

(a) *Negative valency is a loss of* cue. A monitored circulatory process that involves a loss of value corresponds to negative valency. (Note: this implies that movements of value need not always be monitored.)

(b) *Positive valency is a gain of* cue. A monitored circulatory process that involves a gain in value corresponds to positive valency.

(c) *Intensity is rate of exchange of* cue. The rate of exchange of *cue* between substates corresponds to the quantitative intensity of the valenced state.

(d) *Gain in* cue *is contingent on the achievement of goals.* A gain of *cue* can occur when the achievement of a goal is equivalent to the conditions of satisfaction of a reinforcer.

(e) *Loss of* cue *is contingent on the failure of goals.* A loss of *cue* can occur when the failure of a goal is equivalent to the conditions of satisfaction of a negative reinforcer. (The concept of expected reward may be of use here – it would need to be addressed by a more detailed requirements analysis.)

In other words, certain types of 'feelings' are the self-monitoring of adaptations; that is, the non-intentional component of generic 'happiness' and 'sadness' states includes a movement of internal value, which functions to alter the dispositional ability of substates to buy processing power and determine behaviour. Such a process is self-monitored as a 'brute' feeling because value does not refer, unlike beliefs that can be true or false, or goals that can be achieved or not.

Valenced states are present in agent architectures that attempt to meet a requirement for adaptivity. To meet such a requirement it is possible to evolve or design credit-assignment mechanisms that use a domain-independent representation of utility or value, a kind of internal 'common currency'. Such a representation does not refer (in the way that belief and desire-like substates refer) but is internally relational, and it can be gained or lost depending on whether actions are successful or unsuccessful in leading to rewarding consequences.

To be precise, requirement 6 is both a requirement and empirical claim. It requires that the process of credit assignment be monitored by another system. It also makes the empirical claim that in human architectures the monitoring is such that, in circumstances of attachment and loss (see section 5.1.2), we have introspective, direct knowledge of the process of credit assignment, although this information is just one component of the overall mental state. Valency, defined in section 3.1, is the name for this knowledge, and the monitoring of credit assignment is the architectural process that gives rise to it: Phenomenological and functional analyses have converged. Introducing reified conceptions of 'consciousness' at this point would only serve to mystify the convergence. However, artificial architectures that monitor their credit assignment processes may not have the required higher level design features that would justify the description of the architecture being self-conscious of its valenced states; for example, without mechanisms to map a monitored state to a concept, and the use of that concept in the production of natural language, the architecture would be unable to inform us what kind of valenced state it was in.

The design hypothesis of circulation of value opens up the possibility of architectures that generate valenced states far removed from physiology. High level cognitive processes may be saturated with value, allowing the production of semantic messages (not linked to bodily location, or physiological arousal) coupled with losses or gains in quantitative value.

Relations to information processing theories of emotion

In this section the requirements for a computational libidinal economy are briefly compared with two information processing theories of emotion.

(Oatley & Johnson-Laird, 1985; Oatley, 1992; Johnson-Laird, 1988) present a communicative theory of emotions. They emphasise the communicative role of emotion signals within a cognitive architecture, and the communicative role of emotional expression within a social community.

This comparison will concentrate on the former aspect of their theory.

(Oatley, 1992) states that the central postulate of their theory is that emotions are architectural signals that function to communicate significant junctures of plans:

> Each goal and plan has a monitoring mechanism that evaluates events relevant to it. When a substantial change of probability occurs of achieving an important goal or subgoal, the monitoring mechanism broadcasts to the whole cognitive system a signal that can set it into readiness to respond to this change. Humans experience these signals and the states of readiness they induce as emotions.
>
> K. Oatley, *Best Laid Schemes*, p. 50.

Emotions are held to be states that coordinate 'quasi-autonomous' processes in the nervous system. They are a design solution to problems of the transition between plans in systems with multiple goals. For example, if everything is going to plan the agent will be happy. Happiness corresponds to the situation in which subgoals are being achieved without major problems, or if any problems do occur then they can be solved or modified locally without recourse to global problem solving mechanisms. The emotion signal of 'happiness' ensures the architecture continues with the current plan. Conversely, if a major plan or an important subgoal fails, and there are no local 'patches', an emotion signal of 'sadness' is broadcast that causes a transition from following the plan to either doing nothing or/and instigate search for a new, replacement plan. Similarly, if an active plan is frustrated or prevented from coming to fruition by an environmental obstacle a signal, corresponding to anger, can cause the architecture to 'try harder' or agress, that is devote more resources to the achievement of the goal.

Oatley & Johnson-Laird also rely on the distinction between semantic and control signalling. A control signal is thought to be generated by phylogenetically older machinery and, due to its simplicity, need not be parsed or interpreted. An emotion proper has both control and semantic aspects. For example, sadness has a negative phenomenological tone corresponding to the 'sadness' emotion control signal, but can also be about something and have semantic content, such as being sad at one's poor results in an exam. However, global interruption of processing can occur without any emotional state: the actual scope of the communicative theory may be larger than its intended scope (Beaudoin & Sloman, 1993). Only a subset of significant junctures of plans generate hedonic states.

The requirements for a *cle* enrich this theoretical picture by introducing reinforcement learning. The circulation of value is a pattern of flow of control signals. The signal is 'simple', as it is a non-decomposable quantity (useful design properties of a quantitative representation of value were discussed in section 4 providing new reasons for the existence of 'simple' signals), and coordinates a society of relatively autonomous substates.

The communicative theory explains the hedonic tone of happiness and sadness states by positing basic and irreducible valenced control signals. Control signals differ in valency because they differ in their functional roles (e.g., 'sadness' has a 'terminate or change plan' function, whereas 'happiness' has a 'preserve plan' function). In contrast, the monitoring of circulation of value can generate the negative (sadness) and positive (happiness) control signals of the communicative theory: instead of two signals, there is now one (see requirement 6). This is a more parsimonious state of affairs. More importantly, there is a new, previously unidentified functional role for the control signal: the circulation of value implements a type of adaptation. This is not inductive learning of new hypotheses about a domain, but an ordering and reordering of the utility of motivational substates to dispositionally determine behaviour. The single control 'signal' is a quantitative representation of value that can be used to allocate processing resources. It can be both stored and exchanged ('communicated'). On this view, the communication of significant junctures of attachment plans is a secondary or derived functional role of the happiness and sadness control signals: monitoring of circulation can provide this information but the primary purpose of the control signal is to perform credit assignment.

However, the communicative theory accounts for other types of emotional states, whereas the *cle* only accounts for simple affect, that is forms of achievement pleasure and failure unpleasure during processes of attachment. The *cle* in its current form does not account for shorter-term control signals involved in initiating, preserving or terminating action tendencies (Frijda, 1986).

Conceptually, the *cle* can be straightforwardly integrated with the 'attention filter' theory. A subset of perturbant states are valenced perturbances that involve the production of an attention disrupting motivator coupled with the self-monitoring of an exchange of value performing credit assignment (see following section). The possession of *cue* by a substate is linked to its dispositional ability to buy processing power and construct motivators for attentive processing. Due to the high evolutionary importance of developing attachments, libidinal generactivators are likely to be high in value and therefore will tend to commandeer attentive resources and preattentive processing (see requirement 3c). Thoughts pertaining to attachments are likely to be frequent and disrupting.

In the next section it is shown how the *cle* can account for many of the features of attachment and loss.

An example: loss

In colloquial terms, attachment is the process of learning to love someone and loss is the process of learning to live without a loved one when they die, or leave. These are complex processes, involving many more factors than information processing in a cognitive architecture. Also, such events never take exactly the same course, and vary from relationship to relationship. However, the task of this paper is to explain only a subset of the phenomena that generally occur in attachment and loss scenarios, namely positive and negative perturbant states, in particular the intense mental pain that is often associated with grieving.

In (Wright, Sloman & Beaudoin, 1996) we described the kinds of diverse effects that interaction with another person can have upon an architecture, including changing aspects of perceptual and belief systems, the creation of new motive generactivators, motive comparators, joint plans, and predictive models of the other person. These changes were collected under the term ‘attachment structure’, which is a distributed collection of information stores and active components embedded in different parts of the architecture and linked to many other substates.

The *cle* adds to this picture. If interaction with another person satisfies the preconditions of various untaught and learnt reinforcers the *cle* will assign credit to the responsible motivational substates (requirement 1). This is positive feedback during the selective cycle (req. 1d): existing substates determine behaviour that is rewarding, leading to an increase in their *cue* (req. 4b), which, in turn, allows rewarded substates to buy more processing resources and have greater causal powers to dispositionally determine motive management processes (req. 3), and hence behaviour. The process will include the construction of libidinal generactivators (req. 1e) that are specific to the particular person concerned. The generactivators will accumulate *cue* if they produce motives that lead to successful outcomes. As substates gain in value they gain in the ability to commandeer libidinal resources, including forming links with producer substates, or ‘employing’ substates directly. A new ‘branch of production’ will appear, concerned with generating motives pertaining to the person, ensuring the attachment process continues, and various threat and opportunity detectors. In consequence, the attachment process will include moments of (libidinal) goal achievement or failure linked with the monitoring of credit assignment. Such a process will be characterised by valenced states (req. 6). Of course, there will be other types of emotional state apart from valenced states: for example, there may be moments of disappointment, where an expected reward did not occur. The subject, therefore, over a certain period of time, enjoys rewarding interaction with another person who becomes an object of affection. The attachment structure, including the libidinal substates, mediates the mature relationship.

When a loved one dies, particularly if the loved one is a long term partner, a process of grief generally ensues. Grief is not like disappointment, but is more profound, often involving moments of intense mental pain that swamp attention and lead to bouts of crying and howling. Over time these moments become less and less frequent, unless the grief becomes pathological. The *cle* can begin to explain the nature of such painful and insistent thoughts.

The attachment structure is no longer adapted to its environment, and is no longer useful for mediating behaviour. A whole set of concerns is violated; for example, libidinal substates that detect threats to attachment plans will be activated, generating motives for attentive processing. The construction of a motive (semantic signalling) also involves the exchange of an amount of *cue* (control signalling) (req. 3b). If a motivator pertaining to the dead person is generated by libidinal generactivators it becomes a candidate for surfacing. Due to the prior accumulation of *cue* by the libidinal generactivator the motivator will have a high probability of surfacing (req. 3c). If it surfaces management processes will 'decide' (see section 2.2) that it is unsatisfiable, and reject it. The motive will not be 'bought' by management processes, that is the rejected motive will never lead to rewarding consequences. Therefore, the libidinal generactivator expends *cue* (req. 3b) to produce a motive but will never receive *cue* for its product. Due to the difference between preattentive and attentive importance calculations (req. 3c) the production of such motivators is likely to lead to perturbant states, which are emergent states involving a motivator that continues to surface despire being rejected by management processing: in such a state, attention is partially or wholly 'out of control'. This is a crisis of overproduction, in which libidinal generactivators construct unsatisfiable motives that are not bought by other substates. In addition, the death of the loved one may satisfy conditions of satisfaction of negative reinforcers, involving credit assignment processes that reduce the amount of *cue* held by substates. In summary, libidinal substates will gradually lose their accumulated *cue*.

The monitoring of these processes will detect net losses of *cue*, generating mental states of negative valency, which are quantitatively varying, non-intentional states of cognitive unpleasure (req. 6). Therefore, the economic design hypothesis, which takes the form of exchange of *cue* within a *cle*, can ground the folk psychology intuition that some emotions involve a 'release': the accumulated value of libidinal substates is gradually expended in a hopeless attempt to satisfy libidinal reinforcers. This 'flow of value' continues until the *cue* has been 'released' and the substates lose their causal powers. Negatively valenced perturbant states ('painful and insistent thoughts') may be a necessary consequence of adaptive change: the libidinal part of the structure of attachment, no longer useful for motive generation, loses its ability to buy processing power, grab attentive resources and dispositionally determine behaviour. The process of libidinal generactivators expending value to produce unsatisfiable motivators is self-monitored as unpleasurable. This also provides one possible reason why grief differs from disappointment. The former case involves a loss of *cue* that has actually been gained in the past, whereas the latter case involves a failure to gain a reward or achieve a goal predicted to occur in the future.

An important, open question is why higher level management processes cannot simply alter the amount of *cue* possessed by libidinal substates. In this way, substates' ability to buy processing power could be negated by an attentive operation. An architecture that had this ability could avoid the process of a gradual loss of *cue*. Explaining this is outside the scope of the paper, but see (Wright, Sloman & Beaudoin, 1996) for some possibilities.

However, this brief and incomplete consideration of grief demonstrates that the design requirements for a *cle* generate explanations that are psychologically plausible and can begin to unify such phenomena as motivation, reinforcement, adaptive change, and mental pleasure and unpleasure.

Corollary: a reappraisal of libido theory

Freud was concerned with motivational and dynamic aspects of cognition. In this section, the requirements for a computational libidinal economy are compared to aspects of Freudian metapsychology, in particular, his much criticised concept of libidinal energy. Obviously, there is not the space for a full and detailed comparison of the two theories here.

Freudian libidinal economy

Freud held that the instincts are the source of *psychical energy*, which he calls *libido*, *interest*, or *cathectic energy*. Libidinal energy derives from the sexual instincts and is a particular type of the more general cathectic energy.

The *Id* is a subsystem of the mind, consisting of unconscious processes. It is where instincts can attach libido to various *objects*. Object is a technical term, but can be best understood by translating it into modern terminology: it is a cognitive representation of some sort. For example, an instinct may attach libidinal energy to the representation (or object) of another human being, such as a caregiver in early life.

Mental processes, in particular those in the unconscious Id, are regulated by the *pleasure principle*. This is the seemingly simple postulate that psychic processes strive towards gaining pleasure and avoiding unpleasure. The instincts bestow or withdraw libidinal energy to and from various mental objects according to the pleasure principle. For example, a child will discover objects or events in the environment that are associated with pleasurable occurrences. Such objects will have libidinal energy transferred to them, a process sometimes called *cathexis*. It is the investing of energy in the object of desire – an assignment of positive value to it.

Freud held that there is no negation, or contradiction, in the Id. This means that the various unconscious processes are unaware of each other, are entirely selfish, and strive for individual satisfaction. In modern terminology we might say that processes within the Id are relatively autonomous, operate in parallel, and act with mainly local knowledge, unaware of the possible contradictory demands they make on higher level systems.

When discussing libido Freud uses a number of analogies. An 'energy' metaphor is often used, implying that a mental object contains libidinal energy in a latent state, ready to be utilised at any time. A 'hydraulic' metaphor is used, particularly when discussing the dynamic flows of libido within the Id. When discussing the 'striving' nature of instinctual forces, in particular their efforts to circumvent conscious repression, he favours an 'amoeba' analogy. The pseudopodia of the amoeba are the instinctual flows of libido testing out mental pathways in a continual search for satisfaction.

The major functional role of libido is motivational: it is the carrier of instinctual demands. Libido is motivational energy in the sense that it is a 'force' or 'interest' that can direct thought and behaviour. Objects with high libidinal energy *tend* to occupy attention (the Id, for example, places instinctual demands on conscious processes by cathecting libido to various objects). Freud writes –

> We have defined the concept of 'libido' as a quantitatively variable force which could serve as a measure of processes and transformations occurring in the field of sexual excitation.
>
> S. Freud, *Three Essays on the Theory of Sexuality* (1905), p. 138 of (Freud, 1987).

Therefore, libido is also quantitative; it becomes attached to objects in definite amounts. In *Repression* (1915) Freud writes –

> Clinical observation now obliges us to divide up what we have hitherto regarded as a single entity; for it shows us that besides the idea, some other element representing the instinct has to be taken into account, and that this other element undergoes vicissitudes of repression that may be quite different from those undergone by the idea. For this other element of the psychical representative the term *quota of affect* has been generally adopted. It corresponds to the instinct in so far as the latter has become detached from the idea and finds expression, proportionate in its quantity, in processes which are sensed as affects.
>
> S. Freud, *Repression* (1915), p. 152 of (Freud, 1991).

It is unclear what is the precise meaning of affect is in this context. However, if it is taken to mean 'feelings' of whatever kind, as opposed to objects or ideas that have explicit representational content, then libido can be related to non-intentional phenomena (but it should be noted that the conceptual relations between affect, emotions, 'discharge' and libido in Freudian theory is complex). Therefore, the link to affect can be considered a 'weak' property of libido.

There are a number of problems with libido theory. Two major difficulties are the dynamism of libido and its relation to affect. Freud writes –

> ... the mechanisms of repression [conscious or ego-based suppression of motives] have at least this one thing in common: *a withdrawal of the cathexis of energy* (or of *libido*, where we are dealing with sexual instincts).
>
> S. Freud, *Repression* (1915), p. 154-5 of (Freud, 1991).

This quotation (and others) are often ambiguous as to what causes change. Is it libido that is dynamic, withdrawing from an object when it encounters repression, or do the mechanisms of repression direct the withdrawal of libido? Libido itself could be dynamic or the processes that operate upon it. The metaphors chosen by Freud (hydraulics, energy and pseudopodia) imply that it is libido itself that is the agency of change.

Also, Freud did not develop a comprehensive theory of affect or emotion. In *Beyond the Pleasure Principle* (1920) Freud writes -

> Here might be the starting point for fresh investigations. Our consciousness communicates to us feelings from within not only of pleasure and unpleasure but also of a peculiar tension which in its turn can be either pleasurable or unpleasurable. Should the difference between these feelings enable us to distinguish between bound [static, cathected to an object] and unbound [dynamic, flowing between objects] processes of energy? or is the feeling of tension to be related to the absolute magnitude, or perhaps to the level, of the cathexis, while the pleasure and unpleasure series indicates a change in the magnitude of the cathexis *within a given unit of time*?
>
> S. Freud, *Beyond the Pleasure Principle* (1920), p. 337 of (Freud, 1991).

This quotation conflates a number of phenomenological isses. The huge diversity of phenomenological phenomena that can be classified as either ‘pleasurable’ or ‘unpleasurable’ are treated together, without an attempt to distinguish cases. Also, the feelings involved in desiring (‘tension’), which can be either pleasurable or unpleasurable (compare a strong desire to urinate with the pleasure of desire during the sexual act), are considered together with feelings occurrent on the achievement or failure of important longer-term goals (compare winning an olympic gold medal to losing a loved one)[7]. Such (poorly understood) phenomena are then linked with ‘flows’ of cathectic energy, and by extension, libido, without an increase in clarity.

The next section compares libido theory with the computational libidinal economy outlined in this document.

A comparison of the conative universal equivalent and libido

The properties of libido in Freudian metapsychology can be summarised as follows:

1. Dynamic: libido is generally a cause of change, bestowing and withdrawing its attachments (cathexes) to and from various mental objects according to the pleasure principle. It is ‘hydraulic’ in the sense that it flows between mental objects, an ‘energy’ in the sense that it is the ‘fuel’ that causes mental events to occur (such as directing attention and ‘interest’), and ‘striving’ in the sense that it is the active representative of instincts (particularly sexual), seeking to achieve their conditions of satisfaction.

2. Quantitative: it flows and becomes attached to objects in definite quantities.

3. Attentional: libido is correlated with the direction of attention, for example an object with high libidinal energy is likely to gain conscious attention (unless repressed).

4. Motivational: libido is the psychic representative of (strictly speaking, sexual) organic drives or instincts, and represents their motivational ‘push’ in mental life. Attention and motivation are implicitly

[7] This roughly corresponds to the distinction between short-term (preserve vs. terminate) and long-term (positive or negative reinforcement) control states made in (Wright, Sloman & Beaudoin, 1996).

linked in Freudian metapsychology.

5. Non-intentional: libido is to be contrasted with objects that represent things. It does not refer.

6. Unclear relations with affect: the relations between libido theory and 'feelings' or simple affects, such as cognitive pleasure and unpleasure, are unclear.

7. Basic: or primitive, in the sense that the sources of libidinal energy are the instincts, which are deemed to be the innate representatives of our evolutionary heritage.

8. Adaptive: libido flows to and from mental objects according to the pleasure principle, that is libidinal energy transfers to those objects (and their associations) that are linked with positive outcomes, which are conditions that satisfy instincts.

9. Storable: libido is cathected to objects, where it can attached, stored or connected. It is distributed over objects in the Id.

The properties of the conative universal equivalent that circulates within a libidinal economy can be summarised as follows.

1. Passive: in contrast to libido, *cue* is passive and operated upon. The main operator on *cue* is exchange. *Cue* is not hydraulic, 'energising' or striving: the dynamism of the system is within motivational substates, not within the *cue* that is exchanged. However, *cue* does 'flow' within the system, via local exchanges. Also, *cue* is computational or information-theoretic, in the sense that it is a control signal within an information processing architecture. This avoids the vitalistic connotations of Freud's metaphor of 'energy'[8]

2. Quantitative: *cue* is quantitative; it is exchanged and stored in definite quantities.

3. Ability to buy processing power: the possession of *cue* by a substate is a dispositional ability to buy processing power. This includes the ability to construct motivators for attentive processing; hence, *cue* is

[8] However, Freud occasionally used an economic metaphor.

involved in the allocation of attention.

4. Motivational: the sources of *cue* are untaught reinforcement mechanisms, which are examples of *a priori*, evolutionary sources of motivation.

5. Non-intentional: *cue* does not refer, but is relational, specifying a partial ordering (in terms of possession of cue) on system substates. *Cue* is internal economy alone.

6. Clear relations with simple affect: *cue* has well-specified relations to simple affect, to be precise, cognitive valency, which is achievement pleasure or failure unpleasure. The quantitative intensity of valency, and its qualitative differentiation into pleasure or unpleasure is equated with the self-monitoring of credit assignment.

7. Basic: the sources of *cue* are untaught reinforcers, which are held to be genetically specified.

8. Adaptive: exchanges of *cue* perform credit assignment. The *cle* includes untaught reinforcers that reward motivational substates according to the utility of their consequences. This increases their ability to buy processing power and dispositionally determine the behaviour of the system. Over time the system adapts to environmental circumstances. A net gain in value is linked to the achievement of basic goals and can be monitored as positive valency. This process is an example of the pleasure principle, that is *cue* is transferred to those 'objects' that satisfy instincts.

9. Storable: *cue* is stored with substates, and is distributed over the substates of the libidinal economy.

In summary, the economic design hypothesis and related requirements share a number of important features with libido theory. However, there are differences, especially properties 1, 3 and 6. Freud believed that 'psychic energy' was a ubiquitous causal factor in psychological processes, whereas the *cle* is narrower in scope: it is restricted to processes of attachment characterised by valenced states, where valency is a technical concept and is to be distinguished from other forms of pleasure and unpleasure. The theory has been motivated by different concerns and arguments (particularly the emphasis on requirements and design). Also, the *cle* is better specified than libido theory, to the extent that it will be possible to develop simple implementations to help clarify the requirements. Some existing architectures already operate on principles that conform to the *cle*, albeit in a simplified manner, for example Holland's classifier system. This is why the requirements outlined in this document are called requirements for a computational libidinal economy: the name is intended to reflect a convergence of ideas from different perspectives.

CONCLUSION

This paper has outlined basic and preliminary design requirements for a computational libidinal economy that accounts for some types of motivation, learning and affect, in particular valency, a type of cognitive pleasure and unpleasure. The requirements have antecedents in modern information processing theories of emotion and reinforcement learning algorithms, and are strongly related to aspects of Freudian metapsychology, in particular libido theory and cathexis. The economic design hypothesis, abstracted from adaptive systems that use a quantitative, universal representation of value, a kind of internal 'common currency', provides a new, computational version of 'libidinal energy' and avoids the vitalist connotations of Freud's original theory. A claim is that the requirements are in principle implementable, and that existing architectures already exhibit some of the necessary features.

It is admitted that more design work is required before the theory is specified in sufficient detail as to be implementable in a simplified domain. Many of the terms used in this paper need to be made more precise by grounding definitions in a working architecture. An area for further research is to develop agent architectures that integrate reinforcement learning processes with more complex forms of motive management. As an example, (Benson & Nilsson, 1995) describe an implemented agent architecture that attends to multiple, competing goals, selecting between them on the basis of expected reward. Work of this nature will greatly clarify and enrich theoretical concepts, and pose new design problems. Requirements analysis is the first stage of the design based approach and more iterations are required before designs are detailed enough to make testable predictions, or exhibit the kinds of capabilities of human motivational systems.

ACKNOWLEDGEMENTS

My thanks to Aaron Sloman whose extensive comments and criticisms of an earlier version of this paper did much to improve and clarify its content. Also, my thanks to Chris Complin for stimulating discussions and introducing me to classifier systems.

REFERENCES

Bates, J., Loyall, A. B., & Reilly, W. S. (1991). Broad agents. In *Paper presented at AAAI spring symposium on integrated intelligent architectures.* (Available in SIGART BULLETIN, 2(4), Aug. 1991, pp. 38–40).

Beaudoin, L. P. (1994). *Goal processing in autonomous agents.* PhD thesis, School of Computer Science, The University of Birmingham.

Beaudoin, L. P. & Sloman, A. (1993). A study of motive processing and attention. In A.Sloman, D.Hogg, G.Humphreys, Partridge, D., & Ramsay, A. (Eds.), *Prospects for Artificial Intelligence*, pages 229–238. Amsterdam: IOS Press.

Beer, R. D., Chiel, H. J., & Sterling, L. S. (1990). A biological perspective on autonomous agent design. In Maes, P. (Ed.), *Designing autonomous agents: Theory and practice from biology to engineering and back*, pages 169–186. Amsterdam: Elsevier Science Publishers.

Benson, S. & Nilsson, N. J. (1995). Reacting, planning, and learning in an autonomous agent. In Furukawa, K., Michie, D., & Muggleton, S. (Eds.), *Machine Intelligence 14.* Oxford: The Clarendon Press.

Bowlby, J. (1979). *The Making and Breaking of Affectional Bonds.* Tavistock Publications.

Bowlby, J. (1988). *A Secure Base.* Routledge.

Braitenburg, V. (1984). *Vehicles, experiments in synthetic psychology.* MIT Press.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press.

Cichosz, P. (1994). Reinforcement learning algorithms based on the methods of temporal differences. Master's thesis, Institute of Computer Science, Warsaw University of Technology.

Dennett, D. C. (1996). Do animals have beliefs? In Roitblat, H. L. & Meyer, J.-A. (Eds.), *Comparative approaches to cognitive science*, pages 111–118. Cambridge, Massachusetts: “A Bradford Book” The MIT Press.

Donnart, J. Y. & Meyer, J. A. (1994). A hierarchical classifier system implementing a motivationally autonomous animat. In *From Animals to Animats III, Proceedings of the Third International Conference on the Simulation of Adaptive Behavior*. MIT Press.

Dorigo, M. & Colombetti, M. (1993). Robot shaping: developing situated agents through learning. Technical Report TR-92-040, International Computer Science Institute, Berkeley, CA. Revised version.

Doyle, J. (1994). A reasoning economy for planning and replanning. In *Technical papers of the ARPA Planning Initiative Workshop.*

Dyer, M. G. (1987). Emotions and their computations: Three computer models. *Cognition and Emotion*, 1(3):323–347.

Freud, S. (1987). *On sexuality: three essays on the theory of sexuality and other works*, volume VII of *The Pelican Freud Library.* Penguin Books.

Freud, S. (1991). *On metapychology: the theory of psychoanalysis*, volume III of *The Penguin Freud Library.* Penguin Books.

Frijda, N. H. (1986). *The Emotions.* Cambridge: Cambridge University Press.

Frijda, N. H. & Swagerman, J. (1987). Can computers feel? theory and design of an emotional system. *Cognition and Emotion*, 1:235–257.

Green, O. H. (1992). *The Emotions, a philosophical theory.* Kluwer Academic Publishers.

Hayes-Roth, B. (1990). Architectural foundations for real-time performance in intelligent agents. *Journal of Real-Time Systems*, 2:99–125.

Holland, J. H. (1975). *Adaption in natural and artificial systems.* The MIT Press.

Holland, J. H. (1995). *Hidden Order, how adaptation builds complexity.* Helix Books.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: processes of inference, learning and discovery.* The MIT Press.

Johnson-Laird, P. N. (1988). *The Computer and the Mind, an introduction to cognitive science.* Fontana Press.

Kaebling, L. P., Littman, M. L., & Moore, A. W. (1995). Reinforcement learning: a survey. In *Practice and Future of Autonomous Agents, volume 1.*

Maes, P. (1990). Guest editorial: designing autonomous agents. In Maes, P. (Ed.), *Designing Autonomous Agents: theory and practice from biology to engineering and back*, pages 1–2. Amsterdam: Elsevier Science Publishers.

Marx, K. (1970). *Capital, a critical analysis of capitalist production*, volume 1. Lawrence and Wishart. Originally published in 1887.

McCarthy, J. (1979). Ascribing mental qualities to machines. In Ringle, M. (Ed.), *Philosophical Perspectives in Artificial Intelligence.* Sussex: Harvester Press.

Minsky, M. L. (1987). *The Society of Mind.* London: William Heinemann Ltd.

Moffat, D. & Frijda, N. H. (1995). Where there's a will there's an agent. In Wooldridge, M. & Jennings, N. (Eds.), *Intelligent Agents.* Berlin: Springer-Verlag.

Newell, A. (1990). *Unified Theories of Cognition.* Cambridge, MA: Harvard University Press.

Oatley, K. (1992). *Best Laid Schemes.* Studies in Emotion and Social Interaction. Cambridge: Cambridge University Press.

Oatley, K. & Johnson-Laird, P. N. (1985). Sketch for a cognitive theory of emotions. Technical Report CSRP 045, School of Cognitive Science, University of Sussex.

Pepper, S. C. (1958). *The Sources of Value.* University of California Press.

Pfeifer, R. (1994). The 'fungus eater approach' to emotion: a view from artificial intelligence. Cognitive Studies*: Bulletin of the Japanese Cognitive Science Society*, 1(2):42–57. Extended and revised version of an invited talk at *AISB-91*, Leeds, UK. Also available as a technical report from AI Lab, Institute for Informatics, University of Zurich-Irchel.

Powers, W. T. (1988). *Living Control Systems, selected papers of William T. Powers.* Kentucky: The Control Systems Group.

Riolo, R. L. (1988). *A package of domain independent subroutines for implementing classifier systems in arbitrary, user-defined environments.* Logic of Computers Group, Division of Computer Science and Engineering, University of Michigan.

Ryle, G. (1949). *The Concept of Mind.* Hutchinson.

Simon, H. A. (1967). Motivational and emotional controls of cognition. Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979.

Simon, H. A. (1981). *The Sciences of the Artificial* (second ed.). The MIT Press.

Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind.* Hassocks, Sussex: Harvester Press (and Humanities Press).

Sloman, A. (1985). Real time multiple-motive expert systems. In Merry, M. (Ed.), *Expert Systems 85*, pages 1–13. Cambridge: Cambridge University Press.

Sloman, A. (1987). Motives mechanisms and emotions. *Emotion and Cognition*, 1(3):217–234. Reprinted in M.A.Boden (ed), *The Philosophy of Artificial Intelligence*, OUP, 1990.

Sloman, A. (1992). Prolegomena to a theory of communication and affect. In Ortony, A., Slack, J., & Stock, O. (Eds.), *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, pages 229–260. Heidelberg, Germany: Springer.

Sloman, A. (1993a). The mind as a control system. In Hookway, C. & Peterson, D. (Eds.), *Philosophy and the Cognitive Sciences*, pages 69–110. Cambridge University Press.

Sloman, A. (1993b). Prospects for ai as the general science of intelligence. In A.Sloman, D.Hogg, G.Humphreys, Partridge, D., & Ramsay, A. (Eds.), *Prospects for Artificial Intelligence*, pages 1–10. Amsterdam: IOS Press.

Sloman, A. (1995a). Exploring design space & niche space. In *Proc. 5th Scandinavian Conf. on AI, Trondheim*, Amsterdam. IOS Press.

Sloman, A. (1995b). Musings on the roles of logical and non-logical representations in intelligence. In Glasgow, J. & Hari Narayanan, C. (Eds.), *Diagrammatic Reasoning: Computational and Cognitive Perspectives*. AAAI Press.

Sloman, A. (1996). Towards a general theory of representations. In Peterson, D. (Ed.), *Forms of Representation*. Intellect Books.

Sloman, A., Beaudoin, L. P., & Wright, I. P. (1994). Computational modeling of motive-management processes. In Frijda, N. (Ed.), *Proceedings of the Conference of the International Society for Research in Emotions*, Cambridge. ISRE Publications.

Sloman, A. & Croucher, M. (1981). Why robots will have emotions. In *Proceedings of the Seventh International Joint Conference on Aritificial Intelligence*, Vancouver.

Sloman, A. & Poli, R. (1995). Sim_agent: a toolkit for exploring agent designs. In *ATAL-95, Workshop on Agent Theories, Architectures, and Languages*, IJCAI-95, Montreal.

Steels, L. (1994). The artificial life roots of artificial intelligence. *Artificial Life Journal*, 1(1).

Wellman, M. (1995). Market-oriented programming: some early lessons. In Clearwater, S. (Ed.), *Market-Based Control: A Paradigm for Distributed Resource Allocation.* World Scientific.

Wilson, S. W. (1995). Classifier fitness based on accuracy. *Evolutionary Computation*, 3(2):149–185.

Wilson, S. W. & Goldberg, D. E. (1989). A critical review of classifier systems. In *Proceedings of the Third International Conference on Genetic Algorithms*, pages 244–255, Los Altos, California. Morgan Kaufmann.

Wright, I. P. (1994). An emotional agent: the detection and control of emergent states in autonomous resource-bounded agents. Technical Report RP-94-21, School of Computer Science and Cognitive Science Research Centre.

Wright, I. P. (1996). Reinforcement learning and animat emotions. In *From Animals to Animats IV, Proceedings of the Fourth International Conference on the Simulation of Adaptive Behavior*, Cape Cod, MA. The MIT Press.

Wright, I. P., Sloman, A., & Beaudoin, L. P. (1996). Towards a design based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2).