Losing Control Within the H-Cogaff Architecture

Dean Petters

Abstract The paper describes progress towards producing deep models of emotion. It does this by working towards the development of a single explanatory framework to account for diverse psychological phenomena where control is lost. Emotion theories are reviewed to show how emotions can be described according to various emotional components and emotional phases. When applied in isolation these components or phases can result in shallow models of emotion. The CogAff schema and HCogAff architecture are presented as frameworks which can be used to organise and integrate these various separate ways in which emotion can be described. The examples of losing control discussed in this paper include: short-term emotional interrupts to ongoing processing; experiencing grief after the loss of an attachment figure; longer term emotional strengthening of motivation; using strong emotions to guarantee keeping one's own commitments; Freudian Repression as a defensive loss of access to painful information; and self-deception as a general strategy in deceiving others.

1 Introduction

This papers aims to use an analysis of multiple but related psychological phenomena to work towards developing deeper models of emotion and richer cognitive models in general. These phenomena have in common some lapse or loss of control. Many existing ways to model emotion rely upon shallow models which in turn rely upon easily measurable or reportable behaviours and other phenomena in various test situations, such as brain images or physiological data. As Sloman (2001a) notes: *"A desirable but rarely achieved type of depth in an explanatory theory is having a model which accounts for a wide range of phenomena. One of the reasons for shal-*

Dean Petters
School of Computer Science, University of Birmingham, UK, e-mail: d.d.petters@cs.bham.ac.uk

lowness in psychological theories is consideration of too small a variety of cases." The examples where control is lost which are considered in this paper are a diverse collection that vary in duration from momentary episodes of fear to the development of emotional attachments and experiences of grief which can be greatly extended in time. They vary in intensity from how a school pupil is motivated to greater success in school by a wish for their parents to gain pride in them, to examples of murderous and uncontrolled rage. They also vary from phenomena which are described by a central mechanism, such as self-deception resulting from Freudian Repression, to the multiple distributed mechanisms which von Hippel and Trivers (2011) invoke in the formation of self-deception. These phenomena are integrated by explaining them all in terms of a single information processing architecture. In this integration deeper theoretical architecture-based concepts are introduced such as interrupts or disturbances to processing and changes in the locus of control or access to information are introduced. These concepts are then used to identify subsets of the phenomena related to loss of control. Some phenomena may be best described by central processes within an architecture whereas others might focus on relations with the perceived physical or social environment (Sloman, 2001a).

2 What are emotions?

In 'The Expression of Emotions' (Darwin, 1872) Darwin presented relatively shallow behavioural criteria for emotions based upon the production of facial expressions such as blushing or frowning. This approach to describing typical emotional behaviours has been updated in the work of Ekman (2003). Other relatively shallow measurable or reportable criteria for emotions include: physiological measures such as increased heart rate; the activity of specific regions of the brain; the introspectable and reportable experience of bodily changes or desires, such as wanting to run away, or to hurt someone; the experience of interpreting and labelling of situations which trigger emotions, such as appraising a loud noise as a threat; and typical behavioural responses to emotions such as fighting or running away (Sloman, 2001a). These kinds of shallow criteria for emotion can be organised into four categories that describe an emotion as: (1) a neurally implemented response, (2) with a conscious feeling that possesses sensory qualities, (3) including a cognitive process of interpretation, and (4) resulting in a behavioural response. In his recent book 'What is Emotion?' Jerome Kagan states that the belief that human emotion is constituted of these four categories is a widely held view amongst psychology researchers. However, Kagan notes that states may exist which possess some but not all of these criteria and: *"scholars vary* [...]*in the significance they award to each of these four components"* (Kagan 2007, page 23).

One of the problems with using shallow criteria for emotions based on emotional components is that they underspecify emotional phenomena. Kagan himself notes that the behavioural component is not a strong criterion for confirming an emotion has occurred. As he states, *"an emotion need not be accompanied by any behav-*

ior; that is why the concept of 'emotion regulation' was invented" (Kagan, 2007, page 192). Colombetti and Thompson (2008) put forward a critique of emotion research based on how the cognitive component of emotion is currently understood. They suggest that much research in emotion is based upon processes of cognitive appraisal that treat cognitive and bodily events as separate processes. They argue that this is because the cognitive appraisal component of emotion has been conceptualised as an *"abstract, intellectual, 'heady' process separate from bodily events"* (Colombetti and Thompson, 2008, page 45). So according to this embodied view emotional experience may involve an interpretation of events but this interpretation need not involve processes which are mediated by discrete representations which are acted upon by rule manipulating systems. If Colombetti and Thompson (2008) are correct then much more representationally diverse mechanisms for emotion evaluation and appraisal may need to be considered.

The emotional experience component is also not a clear criterion for validating emotions. Not all agree with Kagan's (2007) view that emotions are by definition conscious. For example, people can be in emotional states like jealousy of which they are not conscious, or in long term emotional states of which they are only intermittently conscious like grief, which continue to predispose individuals to take particular actions even at moments they are not conscious of this emotion. As Sloman (2001a) also notes, people may be unaware of their enjoyment when engrossed in a game of football or watching an opera with attention fully engaged. The requirement for a neural implementation for emotion is challenged by researchers who take a functional view of emotions (Arbib and Fellous, 2004). For example, Evans argues that if we say that emotions have to be mediated by neural structures similar to those found in humans then neither computers nor intelligent aliens with exotic brains could ever to be said to have emotions and that *"this is a very parochial view of emotion"* (Evans 2003, page 102).

What the above discussion of emotional componeDnts has shown is that emotions can exist in the absence of one or more of these components. Instead of defining emotions in these terms we can form an integrative framework that considers for particular emotional states which of the four components detailed above are relevant and how those that are relevant are instantiated. In the remainder of the paper, section 3.1 reviews possible frameworks which might support an integration of the various psychological phenomena related to loss of control being considered in this paper. Section 3.2 showing how the Cogaff schema and HCogAff architecture can incorporate a range of phenomena associated with different kinds of emotion including loss of control. Finally section 4 provides a description of the six examples of loss of control which are the focus of this paper.

3 From shallow to deep models of emotion

3.1 Emotions theories organised by components and phases

This section presents an analysis of diverse emotional episodes which have in common that some element of self-control is lost. An integrated explanatory framework for multiple phenomena should be able to include diverse information structures, mechanisms and processes. Such a broad and rich conceptual toolkit can then support deep models of emotional phenomena. One way to bring together a diversity of approaches is to bring together and combine multiple competing theories of emotion. By aggregating together multiple emotion theories we might hope to gain the breadth and richness required for deep models of self control and the loss of self control. Scherer (2010, pages 10-15) is concerned with assessing the differential utility of emotion theories for dynamic modelling of emotional phenomena. In pursuit of this he provides a review of existing emotion theories by grouping a large number of extant theories into eight major 'families' of emotion theory. In Scherer's approach he characterises emotion theories according to two set of criteria. The first set of criteria are categorising emotion theories according to emotional components. The same four components for emotion are used as described above by Kagan (neural implementation, conscious feelings, cognitive processes, and behavioural responses) with the addition of motivation as a fifth emotional component. The second set of criteria are the seven phases of an emotional episode, which according to Scherer are: low level evaluation; high level evaluation; goal/need priority setting; examining action alternatives; behaviour preparation; behaviour execution; and communication and social sharing. Figure 1 shows how eight classes of emotion theory can be mapped onto the 35 subdivisions of a grid formed by having 5 emotion components as one axis and 7 emotion phases the other axis. The eight 'families' of emotion theory categorised according to these criteria are:

- Adaptational theories view emotion as possessing an important adaptive function and are centred around both low and high level evaluation phases (particularly on fear inducing stimuli) and focus on describing and explaining cognitive and physiological components of emotion. LeDoux's (1996) model of fear is an example of this type of theory.
- Dimensional theories differentiate emotions on their position on two dimensions: pleasantness-unpleasantness and arousal. They focus on the feeling emotional component.
- Appraisal theories focus on the cognitive component of emotion whilst also being linked to physiological, expressive and motivational components. They are centred around the high-level evaluation phase of emotions.
- Motivational theories argue that emotions are derived from evolutionary motivational primitives and whilst focusing across physiological, expressive and motivational components are centred on the goal/need priority setting and examining action alternatives phases of emotion.

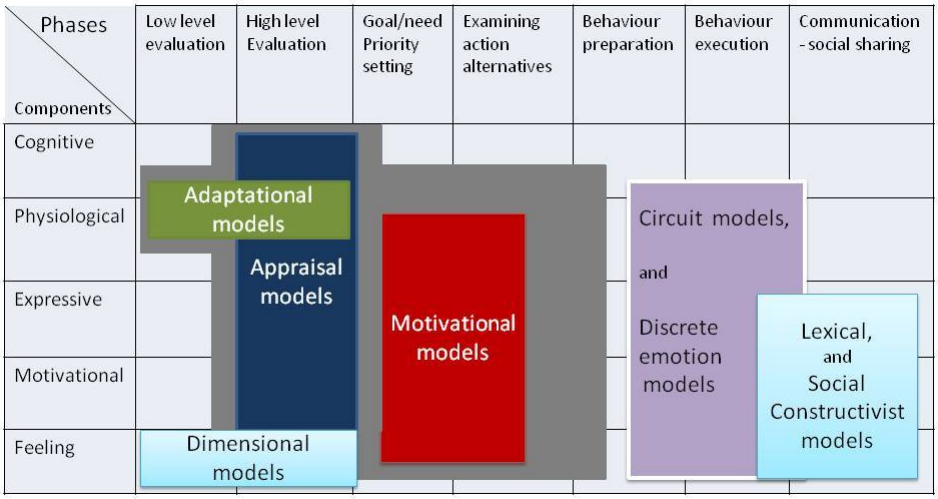


Fig. 1 An adapted version of Scherer's mapping of competing emotion theories onto a grid defined by five emotion components and seven emotional phases. (adapted from Scherer (2010)). This diagram shows that the eight competing emotion theories can be organised into three larger theory-clusters described by Scherer (2010, page 14-15) so that: adaptational, appraisal and motivation models form an 'evaluation and decision' set of theories with their focus on explaining early phases of emotional episodes (this cluster surrounded by a dark grey background); circuit and discrete emotion models together form a 'preparation and execution' set of theories focused on later phases in emotional episodes; and dimensional, lexical, and social constructivist models, form a 'subjective and social' set of theories which are characterised by having less focus on internal 'decision' processes and more on internal subjective feelings and the external social environment.

- Circuit theories focus on expressive and motivational components of emotion and are particularly influenced by pathways and circuits in the brain that are described at a neural level. They explain events in the behaviour preparation and behaviour execution emotional phases.
- Discrete (basic) theories of emotion include ideas from Darwin (1872) and Ekman (2003). They posit that there exist a small number of discrete basic emotions and are focused on physiological, expressive, motivational and feeling components of emotion and are centred around the behaviour preparation and behaviour execution emotional phases.
- Lexical theories assume that the semantic structure of how emotion terms are used will reveal the underlying organisation and determinants of emotions. These theories are focused at the expressive, motivational and feeling components of emotion and are centred around the behaviour execution and communication and social sharing emotional phases.
- Social Constructivist theories are favoured by sociologists and anthropologists who hold that meaning in emotion is socioculturally determined. As with lexical theories, these theories are focused at the expressive, motivational and feel-

ing components of emotion and are centred around the behaviour execution and communication and social sharing emotional phases.

After categorising emotion theories into eight major 'families' Scherer (2010, page 14-15) continues his ambitious classification system by undertaking a further clustering by looking at overlap between these eight families. So figure 1 shows that adaptational, appraisal and motivation models form an 'evaluation and decision' cluster of theories that focus on the initial evaluation, judgement and decision-making phases of an emotional episode; circuit and discrete emotion models together form a 'preparation and execution' cluster whose focus starts where the first cluster finishes. The third cluster brings together dimensional, lexical, and social constructivist models. These theories do not have a shared focus based upon a particular phase as dimensional theories focus on early stages of emotional episodes and lexical and social contructivist theories on later written and interpersonal interactions. However, these 'subjective and social' models do focus more on internal subjective feelings and the external social environment.

Scherer suggests that when choosing a theory as a basis for computational modelling of emotion we can integrate the theories serially - by linking up theories in early to late phases of an emotional episode. What this 'chaining' of theories does not give is an overarching theory of all the non-emotional information processing that fills in the gaps in a complete information processing architecture. In assessing Scherer's classification system as a framework for supporting deep explanations of self-control and the loss of control a key question is: does this system provide the kind of 'wrap-around' conceptualisation necessary? What this means is that self-control may not be located in any single component or phase but occurs between and around components and phases. So initial perceptions can affect later behaviour and the converse can occur. For example, behavioural outcomes can trigger self-control of how situations are perceived or judged. An explanatory framework used to support deep models of self-control and the loss of such control therefore needs to include connections and interrelationships between processes occurring across the 35 elements of Scherer's grid in figure 1. These may not occur in a serial 'pipeline' from early to late phases of an emotional episode. The wide-ranging connections between early and late processes and different emotional components can be the basis of emotion regulation and therefore the basis of self-control. Figure 1 highlights that Scherer's three 'super-family' clusters of emotion theory have a gap between 'evaluation and decision' focussed theories and 'preparation and execution' focussed theories. There is not a single 'super-family' of theories which captures interrelationships of self-reflection and control between all emotional phases and all emotional components in a complete and 'wrapped-around' integrated manner. What we therefore require is an approach which does not force us to decide on which phases or components to focus on at the outset but provide a natural coverage of all components, phases, and the possible control processes between these subdivisions. Such a framework would include very long term dispositions such as attachment status as well as shorter term control states like momentary anger and fear (see figure 2). Aggregation of emotion theories does not on its own provide the integrative links we require to model self-control and the loss of control, perhaps because they

understandably focus mainly on emotional phenomena. So non-emotional processes are therefore more peripheral and given less attention in each individual theory. So when aggregation occurs between emotional theories these non-emotional processes are not included and so not available' to provide a 'glue' to stick all the emotional pieces together. The next section reviews the Cognition and Affect approach which situates emotions within a broader information processing framework where non-emotional processes are considered alongside affective and emotional processes.

LONGER TERM	INTERMEDIATE	SHORTER TERM
Personality, Temperament, Attitudes, Skills, Emotions such as love, grief, Attachment style	**Moods, Beliefs, Preferences, Emotions such as joy, fear, Intentions, Plans, Desires**	**neural and physical events,**

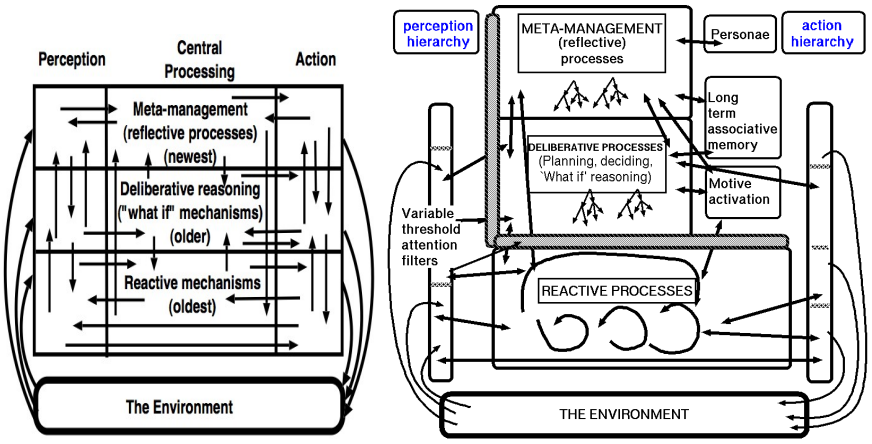
Fig. 2 Classes of semantic control state, which are compared with respect to the approximate duration that each class of control state may exist as a disposition within an architecture (Adapted from Sloman (1995) and Petters (2006)).

3.2 Emotions within the CogAff Schema and HCogAff Architecture

Sloman (2001a) portrays the varying approaches in how emotion is researched as similar to the contradictory opinions expressed by the proverbial ten blind men each trying to say what an elephant is on the basis of feeling only a small part of it. Instead of arguing over which description is right, the ideal solution is for the men describing an elephant to attempt to describe the whole elephant. A way to bring about theoretical integration in emotion research and work towards describing 'the whole elephant' is to form computational models which are based upon cognitive architectures which possess a rich internal structure and show how different emotional classes and phenomena can be produced.

Sloman (2000, 2001a, 2002) has set out the CogAff schema (figure 3(a)) and the HCogAff architecture (figure 3(b)) as conceptual tools which facilitate integrating diverse information processing structures, and mechanisms, and explanations of behavioural phenomena. The CogAff schema is a systematic architectural framework which is intended only as a first approximation to summarising layers of control and cognition produced by evolution. Although there is as yet no agreed conceptual framework for describing architectures, the CogAff schema makes some high level distinctions (Sloman, 2001a). The CogAff schema (figure 3(a)) organises in-

formation processing by overlaying three layers (reactive; deliberative; and meta-management) and three columns (perception, central processing and action). The HCogAff architecture is a special case (or subclass) of CogAff which has not been implemented in its entirety though the production of some subsets have been accomplished (Petters, 2006; Wright, 1997). A very important distinction in the HCogAff architecture is between non-attentive reactive or perceptual processes and the attentive processes which occur within the variable attention filter in figure 3(b). Motive generactivators operate in parallel in the non-attentive reactive component of HCogAff and act to generate to activate motives. They are triggered by internal and external events. They can be thought of as 'scouring' the world for their firing conditions (Wright et al, 1996). When these conditions are met a motivator is constructed which may 'surface' above the attentional filter and be operated upon by processes in the deliberative or meta-management levels. Amongst the processes generated by motivators are: evaluations, prioritisation, selection, expansion into plans, plan execution, plan suspension, prediction, and conflict detection. Management processes can form and then operate upon explicit representations of options before selecting options for further deliberation (Wright et al, 1996). Representations in the deliberative layer can be formed with compositional semantics that allow a first order ontology to be formed which includes future possible actions. The meta level can operate using meta-semantics which allows a second order ontology to be formed which refers to the first order ontology used within the deliberative level (Sloman and Chrisley, 2005).



(a) The CogAff schema (b) The HCogAff architecture

Fig. 3 The CogAff Framework (thanks to Aaron Sloman for permission to use these graphics)

Although the HCogAff architecture has a large scale structure which endures over time, there is constant relocating and transforming of motivators which is termed circulation. As Wright et al (1996) notes, useful control states become more

influential and 'percolate' up a hierarchy of dispositional control states. Ineffective motivators wither away in influence. One important process is 'diffusion', in which the impact of a major motivator leads it to become gradually distributed in myriad control states which can include new motive generators, plans, preferences, predictive models, reflexes and automatic responses (Wright et al, 1996). Meta-management attempts to influence these numerous processes but some are more controllable than others. In summary, the HCogAff architecture is a control system with a rich collection of control states including a numerous ways in which some processes manage or influence other processes (Wright et al, 1996).

Mapping emotional phases to the HCogAff Architecture

A requirement for a integrative theoretical framework for emotion with comprehensive coverage is that it can represent processes occurring in all the seven emotion phases set out by Scherer (2010). The HCogAff architecture (figure 3(b)) represents all phases of information processing resulting from sensation and perception, to central processing, and then motor control (Sloman, 2000, 2002). It should therefore naturally cover all the seven emotional phases described by Scherer (2010). For example, *Low level evaluation processes* might be mapped to the HCogAff architecture's perceptual subsystem of the reactive layer and *High level evaluation* can be mapped to the perceptual and central processing subsystems of the deliberative layer. *Goal/need priority setting* and *examining action alternatives* might be mapped to central processing and action subsystems of the reactive and deliberative layer. *Behaviour preparation*, *Behaviour execution*, and *Communication and social sharing* might all be mapped to different levels of HCogAff architecture's central processing and action subsystems. Further work is needed to provide more details of the possible mappings, perhaps integrating evidence from brain imaging and neuropsychology with computational modelling and behavioural studies.

Mapping emotional components to the HCogAff architecture

An integrative theoretical framework should also be able to focus on the four emotional components described by Kagan (2007)). The distinction between reactive and deliberative processes can help capture both the neural and cognitive emotional components within a single architecture. The action subsystems at the reactive, deliberative and meta processing levels provide a focus on the behavioural component of an emotion. Of the emotional components presented by Kagan (2007) and Scherer (2010) representing subjective emotional feelings provides a significant challenge for any information processing framework. However, Sloman and Chrisley (2003) argue that an implementation of the HCogAff architecture could develop an inherently private and incommunicable ontology for referring to its own perceptual contents and other internal states. That ontology would be produced by

self-organising classification mechanisms and may explain aspects of qualia relevant to subjective emotional feelings (Sloman, 2010).

4 Six ways to lose control

4.1 Emotions and real-time responses

When faced with a decision it takes time to think through all the evidence and options in a completely rational manner. Sometimes an organism or artefact will not have enough time to make a fully rational decision before disaster is upon it and a less considered but faster decision may have been greatly preferable. The idea of emotions as interrupts to ongoing processing was developed by Simon (1967). The central argument in Simon (1967) is that human information processing must cope with multiple needs in an unpredictable environment. These requirements can be met by two classes of mechanism. Firstly, there are goal terminating mechanisms which permit multiple goals to be processed serially without any one need monopolising the processor. So processing for a particular need is stopped when that need is achieved, when too much time has been spent attempting to achieve the need, or progress towards achieving the need is too slow. Simon shows that these mechanisms allow serial processors to respond to a multiplicity of motives existing within a control hierarchy at the same time without any requirement for special mechanisms that represent affect or emotion. However, these mechanisms are inadequate if the system's processing speed is fixed and there are real-time demands on the system. Then provision must be made for an interrupt system. There are two requirements for an interrupt system: processing towards the main goal must go on continuously in parallel with processing that enables the system to notice when interrupts are required; and when real-time needs of high priority are noticed the noticing program must be capable of setting aside ongoing processing and substituting with a quick response.

LeDoux (1996, page 164) describes in detail the neural basis for fear interruptions operating in rat and human brains in a similar manner to how Simon's description is set out above. LeDoux characterises two different routes to appraisal of possibly threatening objects from the sensory thalamus to the amygdala. These are labelled as a 'low road' and a 'high road'. The low road is a 'quick and dirty' processing pathway that bypasses the cortex. This direct thalamo-amygdala route only provides a crude representation of the threatening object but is very fast. The high road is a slower route to action via the cortex. This route eventually ends up in the amygdala but on the way has allowed cortical processing to provide recognition of the object. The high road can be blocked and overridden in an 'emotional hijacking' when the low road takes control - so that we can start responding to dangerous stimuli before we know what they are. As LeDoux notes, this is very useful in dangerous situations but can result in emotional responses that the person experiencing

them does not understand. This mechanism is also involved in anxiety disorders where stimuli trigger uncontrolled and disproportionate fear responses (LeDoux, 1996, pages 239-242).

The HCogAff architecture (figure 3(b)) shows how losing control can result from primary and secondary emotions (Sloman, 2000, 2002). Primary emotions, such as being startled, terrified, delighted, or the thalamo-amygdal fear mechanism described by LeDoux, can be implemented as global interrupts to processing in the lower reactive layer of the HCogAff architecture. Secondary emotions, such as being anxious, apprehensive or relieved, arise from interruptions to deliberative processing in the middle layer. The changing locus of control between processes in the reactive and deliberative layers of the HCogAff architecture helps explain how automatic and controlled processes compete for control in emotional episodes. The third level of the HCogAff architecture is concerned with managing processes that occur in the lower architectural levels. Disruptions or 'perturbances' to this third level are termed tertiary emotions. These can include lapses in attentional control seen in episodes of extreme anger, grief, or longing when in intense love (Petters et al, 2011). Tertiary emotions arise from disturbances to processing in this higher third reflective layer of cognitive architectures. Sometimes these tertiary emotions are impairments but sometimes they can involve acceleration, redirection, or tighter control that avoids a disaster.

Interrupt mechanisms may help overcome likely design limitations on future robots and because of this their use may be unavoidable. As Sloman and Croucher (1981) predict: *"the need to cope with a changing and partly unpredictable world makes it very likely that any intelligent system with multiple motives and limited powers will have emotions."* (Sloman and Croucher, 1981, page 1). However, this does not mean that the loss of control that comes when emotional interrupts occur is good, it is just the consequences of reacting too slowly which is bad.

4.2 Grief as a loss of attentional control

Attachment relations start early in development and last throughout our lives (Petters and Waters, 2010; Petters et al, 2010). Grief is a response to the loss of someone with whom we are strongly attached. Wright et al (1996) provides of detailed description of how grief as a loss of attentional control can be explained within the HCogAff architecture. In brief, when someone interacts with another with whom they are attached the processes of circulation, percolation and diffusion described above in section 3.2 give rise to a distributed multicomponent attachment structure. So this attached individual will possess information about those they are attached to in their perceptual and belief systems. Many motivators will be formed including those which might activate the goals of proximity to the attachment figure or merely to just think about this person. Some of these will then rise in the hierarchy of dispositional control states. These kinds of motivators will interrelate with other control states in complex ways as they diffuse through the architecture. Grief occurs

in response to loss of the attachment figure because this diverse collection of control states cannot be quickly dismantled and will therefore for some time continue to be triggered and gain attention. As Wright et al (1996) note: *"an attachment structure relating to an individual is a highly distributed collection of information stores and active components embedded in different parts of the architecture and linked to many other potential control states. When an attachment structure concerning individual X exists in an agent, almost any information about X is likely to trigger some internal reaction."* (Wright et al, 1996, page 3). Grief is therefore a tertiary emotion which can endure for lengthy periods as dispositions within the architecture and may be largely outside of awareness.

4.3 Loss of control as a detrimental side-effect when emotions boost motivation

Although he did not use the terminology of contemporary Cognitive Science, Hume highlighted the deep links between cognitive states like belief and other thoughts and affective states like motivation and desire when he proposed that: *"reason alone can never be a motive to any action of the will"* (Hume, 1739, page 413). This quote highlights the truism that motivation and preferences are needed for intelligent thought and action (Sloman, 2004). What contemporary approaches can add to Hume's introduction to the importance of motives in reasoning is the concept of an information processing architecture such as the HCogAff architecture within which these motives are processed. Within such an architecture, motivations and preferences do not have to necessarily lead to a loss of control of behaviour or attention.

An ideal in the exercise of will is to gain and maintain control of self and environment rather than lose control as often happens when emotions take over. So an intelligent agent should be able to possess multiple desires and use processes of reasoning to select between options and try and achieve those desires in a controlled and deliberate manner. However, there are at least two exceptions when losing control may be beneficial. As noted above, emotional interrupts may occur to produce quick responses to real-time needs. In addition, humans are opportunistic and can switch at short notice from working towards long term goals to capture short term opportunities. In the HCogAff architecture motivators compete to become active in the deliberative subsystem of the architecture. Sometimes highly beneficial long term goals could remain active for long periods. However sometimes, especially when potent distractions are present or the variable threshold for motive activation to the deliberative subsystem is low due to fatigue or boredom, beneficial long term goals can be supplanted by shorter term and less beneficial goals. Ainslie (2001) describes how humans discount the value of future rewards according to a hyperbolic curve rather than an exponential curve that would produce consistent choice over time. For example, someone on a diet may have no intention of eating cream cake tomorrow but put a cream cake where it is only seconds away from them eating it and their preference may change. Emotions partly cause this problem by giving rise

to intense short term desires like greed, but they also provide a solution. Short term temptations can be combated by longer term emotions like pride or a strong sense of love or duty.

So emotional 'hijackings' do not just have to be short-term interrupts where control is returned after the moment of danger has passed. They can also last for considerably longer durations to increase long term motivation. Someone who has lost out in some way to another person is more likely to confront them and perhaps gain redress if bolstered with the action readiness and extra motivation provided by being angry. Emotions can involve increased heart rate and preferential blood circulation to the muscles providing increasing action readiness but at the cost of performance in fine grained control. There also is an information level effect where emotions can lead to stronger commitment to goals. For example, because someone is in an emotional state they may be less likely to give up on their goals or consider evidence which might suggest taking an alternative course of action.

The public expression of strong emotions which are linked to greater commitment can also have social benefits, but these benefits might be gained by convincingly portraying these emotions without really experiencing them. Sloman (2000) discusses how a teacher may discover that real anger can be used to control a classroom, and learn to become angry. However, even if the real-anger provides an overall benefit, this loss of control may still lead to some detrimental side-effects. The teacher would be better to express anger convincingly without really being angry with its attendant loss of control.

Passions linked to love, empathy, anger, fear, duty and honour provide motivation and may help overcome distractions like hunger, tiredness and from other less important goals. From an engineering and therapeutic perspective it is clear that many of the motivational benefits of emotion might be provided by non-emotional alternatives. However, in the context of an architecture biased to short term opportunities the extra motivation provided by emotions may provide an important function in delivering long term commitment. Within the HCogAff architecture emotions may not just be triggered when interrupts are needed to cope with fast moving events but also to halt newly surfacing motivators which may distract from ongoing long term goals. The next section will discuss circumstances where loss of control might itself be a benefit rather than merely being a detrimental side-effect arising from otherwise beneficial mechanisms.

4.4 Loss of control which enables total commitment

Humans, more than other animals, can display false emotions. We can also detect such deceptions. Thus the possibility of an evolutionary arms race between deceiver and deceived. Many commentators have suggested a 'Machiavellian Hypothesis' that 'cognitive arms races' between deceiver and deceived may have been a key driving force in the development of human intelligence (Trivers 1971, Humphrey 1976, Alexander 1987, 1990, Rose 1980, Miller 1983, all cited in Pinker (1998)) and

also that of other non-human primates (Byrne and Whiten, 1988; Whiten and Byrne, 1997). If detection of sham anger is likely and so someone cannot convincingly pretend to be angry they may fall back on actually being angry. However, real anger is a dangerous tool which can end badly for the person who has lost control in this way.

Pinker (1998) draws upon the work of a number of researchers (Schelling 1960, Trivers 1971, 1985, Daly and Wilson 1988, Hirshleifer 1987, and Frank 1988, all cited in Pinker 1998) who have all independently proposed an emotional mechanism which Pinker terms the 'Doomsday Machine'. This label is adopted by Pinker from the film Dr Strangelove. In this film the Russians have developed a network of underground bombs with the potential to kill all life on earth. These bombs will be set off automatically in the event of an attack by another country or an attempt to disarm it. The essential property of the Doomsday Machine is that once set up and turned on there is no going back. It makes the Russians immune to threats and blackmail. Pinker quotes the Dr Strangelove character in this film to emphasise how voluntarily losing options is the essence of the Doomsday machine:

> *"But," Muffley said, "is it really possible for it to be triggered automatically and at the same time impossible to untrigger?"*
>
> *... Doctor Strangelove said quickly, "But precisely. Mister President, it is not only possible, it is essential. That is the whole idea of this machine. Deterrence is the art of producing in the enemy the fear to attack. And so because of the automated and irrevocable decision making process which rules out human meddling, the Doomsday Machine is terrifying, simple to understand, and completely credible and convincing."* (quoted in Pinker, 1998, page 408)

Pinker shows that Dr Strangelove's description of how the Doomsday Machine operates in Nuclear Deterrence Strategy can be transferred to human interpersonal conflict. In fact, we can use exactly the same words as Dr Strangelove in describing the effects of the kinds of extreme and uncontrolled rage possessed by individuals who have committed work place massacres in the USA or who run Amok in Indochinese cultural contexts. The triggering of such strong emotions allows such individuals to possess responses which can be 'irrevocable', 'terrifying', and 'simple to understand'. The threat from such individuals is also 'credible and convincing' as long as their emotional state can be distinguished from sham-emotion pretence. Examples of extreme rage are not always entirely non-cognitive processes. As Pinker (1998, page 364) describes, such rampages are preceded by lengthy brooding over failure and involve aspects of planning. They are better interpreted as cognitive strategies where compliance to their commitments has been locked in by strong emotions. However, Pinker goes further and asks whether emotions in general allow individuals to get their own way and enforce their will over others in everyday situations. Pinker (1998) presents a strong view of the intimate interrelation between emotion and reason:

> *"People consumed by pride, love, rage have lost control. They may be irrational. They may act against their interests. They may be deaf to appeals.* [...] *But though this be madness*

> *there be method in it. Precisely these sacrifices of will and reason are effective tactics in the countless bargains, promises, and threats that make up our social relations.*[...]*The passions are no vestige of an animal past, no wellspring of creativity, no enemy of intelligence. The intellect is designed to relinquish control to the passions so that they may serve as guarantors of its offers, promises, threats against suspicions that they are lowballs, double-crosses and bluffs. The apparent firewall between passion and reason is not an ineluctable part of the architecture of the brain; it has been programmed in deliberately, because only if the passions are in control can they be credible guarantors."* (Pinker, 1998, page 412-413)

This a quite a grim view of human relations. Pinker is suggesting that people use their emotions to implicitly threaten others. These threats are different from merely threatening actions like hitting or shouting at people because it is the lack of control that is the real danger. In section 4.2 processes such as circulation, percolation and diffusion were invoked in explaining how attachments form and ultimately result in the loss of control of attentive processes experienced in grief. These same processes can also account for how individuals build up highly distributed collections of motivators that can give rise to Doomsday Machine emotional responses. Over time motivators for gaining revenge, keeping face, or maintaining honour and social status can become embedded in different parts of the architecture. Through diffusion these motivators become linked to and interrelate with many other potential control states in complex ways. So when a moment comes that they are triggered the individual experiencing these active motives will struggle to act in their current best interests if this involves controlling and subduing a large network of other interlinked motivators. If Doomday Machine emotions remain untriggered then the individual may benefit. Or an individual that merely threatens Doomsday responses may bring about adverse reactions and avoidance by others.

4.5 Loss of control - due to blocking of access to painful information

The concept of repression was developed by Freud to explain behavioural patterns observed in his clinical practice in Vienna in the 1890s and early years of the Twentieth Century (Freud, 1925—1995; Storr, 1989). Freud's practice included a large number of individuals complaining of a phenomenon which came to be known as Conversion Hysteria. The reminiscences of hysterics were notable in two regards: they were not easily accessible to conscious recall; and they were often painful, shameful or alarming (Storr, 1989). As Freud described:

> *"How had it come about that the patients had forgotten so many of the facts of their external and internal lives but could nevertheless recollect them if a particular technique was applied? Observation supplied an exhaustive answer to these questions. Everything that had been forgotten had in some way or other been distressing; it had been either alarming or painful or shameful by the standards of the subject's personality. It was impossible not to conclude that that was precisely why it had been forgotten - that is, why it had not remained conscious. In order to make it conscious again in spite of this, it was necessary to overcome*

> *something that fought against one in the patient; it was necessary to make efforts on one's own part so as to urge and compel him to remember. The amount of effort required of the physician varied in different cases; it increased in direct proportion to the difficulty of what had to be remembered. The expenditure of force on the part of the physician was evidently the measure of a* ***resistance*** *on the part of the patient. It was only necessary to translate into words what I myself had observed, and I was in possession of the theory of* ***repression.****"* (Freud, 1925—1995, page 18)

Over many hours of observation Freud came to see the mind as involving conflict between conscious thoughts and unconscious emotions 'trying' to become conscious and be discharged, like a boil trying to reach the surface of the skin and release its toxins (Storr, 1989). However, unlike in the case of a boil, Freud considered that unconscious psychically toxic thoughts were held below consciousness by an active process of defensive control which he termed repression.

For Freud, the main function of repression is minimisation of psychic pain and distress. More recently, from the perspective of contemporary Evolutionary Psychology, Nesse and Lloyd (1992) have proposed that the capacity for repression is an evolutionary adaptation. According to Nesse and Lloyd (1992), some of the adaptive functions which repression may provide include:

- Controlling mental pain in a similar way that endorphins control physical pain. Although sensing pain is usually a useful evolved capacity as the cause of the pain can be removed or avoided, if no actions can be taken to reduce pain and it serves no other continued purpose (such as learning or avoidance) then removing it from awareness may be adaptive as it removes a source of distraction from ongoing thought.
- Inhibiting conscious recognition of socially unacceptable impulses. These impulses may not be intended to be carried out but would rather involve fantasizing. Again, removing them from consciousness would decrease anxiety and may remove a source of distraction.
- Repressing the thought of a friend's selfish motives so that a friendship can be maintained. Nesse and Lloyd (1992) describe how certain schizophrenics have been reported to possess an uncanny ability to apprehend secret and unsavoury motives in others. Nesse and Lloyd speculate that this ability may be due to interference with the normal ability to adaptively deceive oneself about the motives of others. This toleration of selfish motives in one's friends may be adaptive or maladaptive but certainly reduces exposure to painful thoughts.

Nesse and Lloyd also propose that the self-deceptive nature of repression allows people to follow their own selfish motives with less chance of detection by others. This fourth proposal for an adaptive function for repression is less about psychic defense and was inspired by ideas presented in Trivers (1976). The next section presents a detailed account of Triver's updated views on possible functions for self-deception (von Hippel and Trivers, 2011).

Nesse and Lloyd's attempt to link repression with contemporary Evolutionary Psychology is weakened by claims that repression is not a human universal but arises in particular 'rule-bound' cultural environments. The type of case upon which

Freud's theory was originally based - severe conversion hysteria - are seldom observed today. This has been explained by the fact that Freud undertook his studies with subjects from a particular cultural and historical milieu - the Viennese upper or upper middle class culture of the 1890s and early Twentieth Century. Reasons given for the drop in frequency of hysteria include societal emancipation from the constrictions of Victorian culture and changes in emotional literacy in the 20th century. These changes would be expected to decrease the 'need' for repression (Shorter, 1993; Stone et al, 2008). Stone et al (2008) provides a contrasting view that suggests hysteria has not disappeared but rather interest has waned in it by clinicians and patients.

Freud lacked an appropriate theory of information processing and control mechanisms but contemporary frameworks such as the HCogAff architecture may be able to capture the spirit of Freud's conceptual ideas. As figure 3(b) shows the HCogAff architecture can include a number of 'personae' which are high level culturally determined templates which cause global features of the behaviour to change, e.g. switching between bullying and servile behaviour (Sloman, 2001b). Sloman (2010) argues that personae are required because metamanagement subsystems may need different monitoring and control regimes for use in different contexts. Personae might therefore be a mechanism whereby the repression of painful thoughts could be implemented in the HCogAff architecture. Freud noted that repressed thoughts take effort to retrieve. This effort may be the personae being switched, merged or given transparent access to each other.

4.6 Loss of Control - due to biases in what information is processed

Freud's and Nesse and Lloyd's approaches to self-deception are mostly defensive, invoking repression as a process which removes painful thoughts from conscious awareness. In contrast, von Hippel and Trivers (2011) present a more offensive function for self-deception. The central claim of von Hippel and Trivers (2011) is that *"self deception evolved to facilitate interpersonal deception by allowing people to avoid the cues to conscious deception that might reveal deceptive intent"* (von Hippel and Trivers, 2011, page 1). This idea was first presented by Trivers in a foreword to 'The Selfish Gene' (Dawkins, 1976):

> *"If (as Dawkins argues) deceit is fundamental in animal communication, then there must be strong selection to spot deception and this ought, in turn, select for a degree of self-deception, rendering some facts and motives unconscious so as to not betray by the subtle signs of self-knowledge the deception being practiced. Thus, the conventional view that natural selection favours nervous systems which produce ever more accurate images of the world must be a very naive view of mental evolution. "* (Trivers, 1976, pages 19-20)

In their recent treatment, von Hippel and Trivers (2011) expand on Triver's 1976 conjecture and suggest self-deception is a particularly good way to deceive others because it eliminates the costly cognitive load that can be involved in carrying out

deception of others, and it also minimises retribution. Having lots of incorrect beliefs about the world is not beneficial but von Hippel and Trivers suggest that when a person deceives themselves by bolstering their own positive qualities and other peoples negative qualities this can lead to greater confidence and hence social and material advancement.

Where von Hippel and Trivers (2011) differ from the approaches of Freud and Nesse and Lloyd described above is in the variety of processes they claim can give rise to self-deception. What Freud's original approach and Nesse and Lloyd's more recent approach to explaining repression have in common is that the repressed individual possesses two separate representations of reality with truth preferentially stored in the repressed unconscious component and falsehood preferentially in the conscious mind (von Hippel and Trivers, 2011). Von Hippel and Trivers suggest that in addition *"the dissociation between conscious and unconscious memories combines with retrieval-induced forgetting and difficulties distinguishing false memories to enable self-deception by facilitating the presence of deceptive information in conscious memory while retaining accurate information in unconscious memory"* (von Hippel and Trivers, 2011, page 6). They also propose that implicit and explicit attitudes dissociate which enables people to express to others socially desirable attitudes whilst simultaneously acting on socially undesirable attitudes. This is because the socially undesirable attitudes are relatively inaccessible and so help the individual maintain plausible deniability. A further suggested mechanism that helps implement self-deception is a dissociation between automatic and controlled processes. So according to von Hippel and Trivers, whilst controlled processes pursue goals that an individual has deceived themselves that they want to pursue, automatic processes pursue the true but hidden goals. As von Hippel and Trivers note: by *"by causing neurologically intact individuals to split some aspects of their self off from others, these dissociations ensure that people have limited conscious access to the contents of their own mind and to the motives that drive their behaviour. In this manner the mind circumvents the paradox of being both deceiver and deceived."* (von Hippel and Trivers, 2011, page 6)

The processing biases which von Hippel and Trivers (2011, pages 7-11) detail as bringing about a state of self-deception include: (1) biased information search (which can vary in amount of searching; selection of searching and selective attention); (2) biased interpretation; (3) misremembering; (4) rationalisation; and (5) convincing the self that the lie is true. These biases in processing can be situated within the HCogAff architecture. For example, the distinction between reactive and deliberative processes might be related to the distinctions von Hippel and Trivers (2011) discuss between conscious and unconscious memories, implicit and explicit attitudes, and controlled and automatic processes. In addition, processing biases such as biases in information search, biased interpretation and misremembering might all arise from tertiary emotions where meta-management processes are perturbed. In addition to explaining how these biases can form, the HCogAff architecture should also attempt to explain why meta-management processes do not uncover self-deception

5 Conclusion

This paper has shown that the explanatory framework offered by the CogAff schema and HCogaff architecture provides a promising opportunity for producing deep emotion models from an integrative analysis of differing types of loss of control. In particular, primary emotions such as fear or happy surprise and secondary emotions such as worrying about future events can be explained as interrupts to processing in the HCogAff architecture. Processes which describe how motivators relocate and transform within the HCogAff architecture, such as circulation, percolation and diffusion can be used to explain a range of tertiary emotions such as grief and uncontrolled rage. Self-deception involves losing access to 'true' information and how this occurs may spur theoretical developments in the HCogAff architecture. Future work may also not only deepen and enrich a HCogAff explanation of the phenomena discussed above but allow other complex emotional phenomena to be integrated within this explanatory framework.

References

Ainslie G (2001) Breakdown of Will. Cambridge University Press, Cambridge:

Arbib M, Fellous JM (2004) Emotions: from brain to robot. Trends in cognitive sciences 8:554–61

Byrne R, Whiten A (1988) Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans. Oxford University Press, Oxford

Colombetti G, Thompson E (2008) The feeling body: towards an enactive approach to emotion. In: *Developmental Perspectives on Embodiment and Consciousness*, eds. W.F. Overton U. Muller U, & J.L. Newman, Lawrence Erlbaum Ass., New York, pp 45–68

Darwin C (1872) The Expression of the Emotions in Man and Animals. Harper Collins, London, (Reprinted 1998)

Dawkins R (1976) The Sefish Gene. Oxford University Press, Oxford, New York

Ekman P (2003) Emotions Revealed. Times Books, New York

Freud S (1925—1995) An autobiographical study. In: Gay P (ed) The Freud Reader, Vintage, London

von Hippel W, Trivers R (2011) The evolution and psychology of self-deception. Behavioral and Brain Sciences 34:1–56

Hume D (1739) A treatise of human nature (A Critical Edition, David Fate Norton and Mary J. Norton (eds.), Oxford, Clarendon Press, 2007.)

Kagan J (2007) What is Emotion? Yale University Press, New Haven: MA

LeDoux J (1996) The Emotional Brain. Simon & Schuster, New York

Nesse R, Lloyd A (1992) The evolution of psychodynamic mechanisms. In: *The Adapted Mind: Evolution Psychology and the Generation of Culture* , eds. J.H. Barkow and L. Cosmides and J. Tooby, OUP, Oxford, pp 601–624

Petters D (2006) Designing agents to understand infants. PhD thesis, School of Computer Science, The University of Birmingham, (Available online at http://www.cs.bham.ac.uk/research/cogaff/)

Petters D, Waters E (2010) A.I., Attachment Theory, and Simulating Secure Base Behaviour: Dr. Bowlby meet the Reverend Bayes. In: Proceedings of the International Symposium on 'AI-Inspired Biology', AISB Convention 2010, AISB Press, University of Sussex, Brighton, pp 51–58

Petters D, Waters E, Schönbrodt F (2010) Strange carers: Robots as attachment figures and aids to parenting. Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems 11(2):246–252

Petters D, Waters E, Sloman A (2011) Modelling Machines which can Love: From Bowlby's Attachment Control System to Requirements for Romantic Robots. Emotion Researcher 26(2):5–7

Pinker S (1998) How the Mind Works. Penguin Books, London

Scherer K (2010) Emotion and emotional competence: conceptual and theoretical issues for modelling modelling agents. In: Scherer K, Banziger T, Roesch E (eds) Blueprint for affective computing: a sourcebook, OUP, Oxford, pp 3–20

Shorter E (1993) From Paralysis to Fatigue: History of Psychosomatic Illness in the Modern Era. Simon and Schuster, London, UK

Simon HA (1967) Motivational and emotional controls of cognition

Sloman A (1995) What sort of control system is able to have a personality? (Presented at Workshop on Designing personalities for synthetic actors, Vienna, June 1995)

Sloman A (2000) Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In: Dautenhahn K (ed) Human Cognition And Social Agent Technology, Advances in Consciousness Research, John Benjamins, Amsterdam, pp 163–195

Sloman A (2001a) Beyond Shallow Models of Emotion. Cognitive Processing: International Quarterly of Cognitive Science 2(1):177–198

Sloman A (2001b) Evolvable biologically plausible visual architectures. In: Cootes T, Taylor C (eds) Proceedings of British Machine Vision Conference, BMVA, Manchester, pp 313–322

Sloman A (2002) How many separately evolved emotional beasties live within us? In: Trappl R, Petta P, Payr S (eds) Emotions in Humans and Artifacts, MIT Press, Cambridge, MA, pp 29–96

Sloman A (2004) Simulating infant-carer relationship dynamics. In: Proc AAAI Spring Symposium 2004: Architectures for Modeling Emotion - Cross-Disciplinary Foundations, Menlo Park, CA, no. SS-04-02 in AAAI Technical reports, pp 128–134

Sloman A (2010) An Alternative to Working on Machine Consciousness. International Journal of Machine Consciousness 2(1):1–18

Sloman A, Chrisley RL (2003) Virtual machines and consciousness. Journal of Consciousness Studies 10(4-5):113–172

Sloman A, Chrisley RL (2005) More things than are dreamt of in your biology: Information-processing in biologically-inspired robots. Cognitive Systems Research 6(2):145–174

Sloman A, Croucher M (1981) Why robots will have emotions. In: Proc 7th Int. Joint Conference on AI, Vancouver, pp 197–202

Stone J, Hewett R, Carson A, Warlow C, Sharpe M (2008) The 'disappearance' of hysteria: historical mystery or illusion? Journal of The Royal Society of Medicine 101:12–18, 1

Storr A (1989) Freud. OUP, Oxford

Trivers R (1976) Foreword. In: *The Selfish Gene*, eds. R. Dawkins, Oxford University Press, New York, pp 19–20

Whiten A, Byrne R (1997) Machiavellian Intelligence. Vol. 2, Evaluations and Extensions. Cambridge University Press, Cambridge

Wright I (1997) Emotional agents. PhD thesis, School of Computer Science, The University of Birmingham

Wright I, Sloman A, Beaudoin L (1996) Towards a design-based analysis of emotional episodes. Philosophy Psychiatry and Psychology 3(2):101–126