Minds have personalities - Emotion is the core

Darryl N. Davis

Computational Intelligence and Cognition Research Group,

Department of Computer Science, University of Hull,

Kingston-upon-Hull, HU6 7RX, U.K.

D.N.Davis@dcs.hull.ac.uk

Abstract

There are many models of mind, and many different exemplars of agent architectures. Some models of mind map onto computational designs and some agent architectures are capable of supporting different models of mind. Many agent architectures are competency-based designs related to tasks in specific domains (e.g. COG). The more general frameworks (e.g. ACT-R, AIS, SOAR) map across tasks and domains. A number of models for synthetic minds are based on analyses and observations of human minds. These types of agent architectures are capable of performing certain behaviour and cognitive competencies associated with a functioning mind. There is a problem with many of these approaches when they are applied to the design of a mind analogous in type to the human mind – there is no core to mind in any of these theories or designs other than an information processing architecture. As any specific architecture is applied to different domains, the information processing content (knowledge and behaviors) of the architecture changes wholesale. From the perspective of developing intelligent computational systems this is more than acceptable. From the perspective of developing functioning (human-like) minds this is problematic – these models are in effect emotionally autistic. If mind is an ongoing characteristic of an entity of a certain level of complexity and a mind is capable of moving through many different control states, from where do the control patterns that stabilize a mind as an ongoing (developing) personality emanate? Our current work on this theme presents an emotion-based core for mind. This work draws on evidence from neuroscience, philosophy and psychology. As an agent monitors its interactions within itself and relates these to tasks in its external environment, the impetus for change within itself (i.e. a need to learn) is manifested as an unwanted combination of emotions. Such a control state can lead to the generation of internal processes requiring the agent to modify its behavior or processes in some way. The modification of an agent's internal environment is then described in terms of an emotion motivated mapping between its internal and external environments. Cognition and underlying processes are used to navigate the agent-oriented internal environment of emotion. It is suggested that personality traits are a manifestation of this emotion core. Personality becomes an emergent property of the cognitive architecture and its (pre-)disposition to concentrate on certain tasks and favour specific instances of control states. Personality traits affect and influence the different categories of cognitive and animated behavior. Moods arise from the interaction of current temporally-global niche roles (the favouring of certain aspects of emotion space) and temporally-local drives that reflect the current focus of the deliberative processing as perceived by the reflective layer. Temporally-global drives are those associated with the agent's overall purpose related to its current, possible and desired niche spaces. Temporally-local drives are related to ephemeral states or events within the agent's environment or itself. The (high-level) niche-seeking drives (or dispositions) together with the more orthodox control states bind the theoretical model together and allow a synthetic agent to become complete and exhibit a (non-shallow) personality.

Introduction

For much of its history, cognitive science has positioned emotion as the poor relation to cognition. This paper aims to justify a stance on (human-like) minds that places emotion as the core. It is not possible to review all pertinent evidence within the remit of this paper. There is a considerable amount of work from neuroscience on what parts of the central nervous system have a role to play in emotions and relevant aspects will be addressed. Research from psychology, philosophy and psychiatry will be presented. A sketch of a computational theory of mind (primarily from the agent perspective) will be then be considered in the light of this evidence. This leads onto the presentation of preliminary experimental work that models emotions as the core of a computational agent architecture.

Here we completely reject the definition of emotions as "*..examples of non-problem-solving non-behaviour*" (Gunderson 1985:72). Emotion has many functions including the valencing of emerging problems and challenges in terms of emotional intensity and emotion type. Such a function is a precursor to problem solving. The conjecture is that the computational modeling of human-like minds is impossible unless a silicon/digital analog to human-like emotions is possible. Our efforts in producing computational cognition may lead to the development of intelligent problem-solvers of many types, but the simulation of the human mind requires more than intellectual processes. Much of cognitive science and artificial intelligence adopts a modular approach to cognition. If we can solve vision, memory, attention, language, we can build an artificial brain. Where is the glue? When a comprehensive, silicon based model of the human brain is created without emotions, it will be diagnosed as autistic! This approach to cognitive science is one that Harré (1994) regals against - the individual as passive observer of the computational processing that is that person's cognition. To rephrase a previous revolution in artificial intelligence: *human-like intelligence requires embodiment of the supporting computational infrastructure not only in terms of an external environment but also in terms of an internal (emotional) environment.* This paper places emotion at the core of mind.

Psychology and Emotions

The nature of emotions and the relation to thought have been analysed since the dawn of western civilisation. Plato degrades them as distorting rationality. Aristotle denotes long tracts to their categorisation and impact on social life. For Darwin emotions in adult humans are a byproduct of evolutionary history and personal development; serving a minimal function in everyday life. Over the last hundred years of psychology (from James onwards) the study of emotion has waxed and waned with theories of emotion typically rooted in discussions of physiological and non-rational impulses and drives. An exception is the "cognitive" school of emotion dating from Paulhan (1887) through to Schacter and Singer's (1962) influential experiments with adrenaline and the effect of social context on emotive appraisal.

A standard introduction to psychology from the 1970s (Lindsay and Norman 1972) summarise much the experimental work on emotions in suggesting that emotional states are manipulable through cognitive processes (in particular expectations), physiological states and environmental factors. They conclude that cognition (particularly memory, motivation, attention and learning) and emotions are intimately related. In Newell's seminal work on cognition (Newell 1990), emotion is not indexed and is only discussed in any length in relation to social aspects of a cognitive agent in the final chapter. Although Newell acknowledges this fact, it reflects a trend in cognitive science to place emotion as subordinate to rationality and cognition. Despite pointers to the importance of understanding emotion for cognitive science (e.g. Norman 1981), cognitive science all too readily follows as a modern day Stoic successor to Plato in minimising the role of emotion. A leading volume on the dynamics approach to cognition (Port and Van Gelder, 1995) is no exception – particularly odd if emotion is viewed as the flow and change of cognitive predisposition over time and across occasion (Lazurus 1991).

Ortony et al (1988) consider cognition to be the source of emotion, but that unlike many other cognitive, emotions are accompanied by visceral and expressive manifestations. They consider valence (i.e. positive-neutral-negative) and appraisal (cognitive reflection of these valencies) as the primary basis for describing an emotion. They differentiate emotions from non-emotions on the basis of whether a valenced reaction is necessary for that state. Non-emotion states (e.g. abandonment) can give rise to causal chains of emotive reactions leading to highly valenced (emotive) states. They suggest that there are basic classes of emotion related to valenced states focussed on events (pleased vs. displeased), agents (approving vs. disapproving) and objects (liking vs. disliking). Specific emotions are instances and blends of these types and subclasses. Emotions of the same type have eliciting conditions that are structurally related. They

reject the idea of emotions such as anger and fear being fundamental or basic emotions. The cognitive processing that appraises emotions is goal-based and resembles the type of processing and structures discussed in motivation for autonomous agents (e.g. Beaudoin and Sloman 1993, Davis 1996).

Oatley and Jenkins (1996) define emotion as "a state usually caused by an event of importance to the subject. It typically includes (a) a conscious mental state with a recognizable quality of feeling and directed towards some object, (b) a bodily perturbation of some kind, (c) recognizable expressions of the face, tone of voice, and gesture (d) a readiness for certain kinds of action". Similar definitions are given by others (e.g. Frijda 1986). Personality traits lasting years (or a lifetime) are usually tightly bound to qualities of emotions. A number of other psychologists (e.g. Power and Dalgleish 1997) appear to be in agreement in defining what are basic emotions:

- Fear defined as the physical or social threat to self, or a valued role or goal.
- Anger defined as the blocking or frustrations of a role or goal through the perceived actions of another agent.
- Disgust defined as the elimination or distancing from person, object, or idea repulsive to self and to valued roles and goals.
- Sadness defined as the loss or failure (actual or possible) of valued role or goal.
- Happiness defined as the successful move towards or completion of a valued role or goal.

It is suggested that these five suffice as the basic emotions as they are physiologically, expressively and semantically distinct plus they have a biological basis. There are cases for other emotions to be considered as further basic emotions. From a perspective of classifying emotions using different and distinctive universal signals (Ekman & Davidson 1994) surprise is included in this fundamental set. However, from the perspective of classifying emotions based on distinctive physiological signs (see Power and Dalgleish 1997), the basic set is reduced to fear, anger, disgust and sadness. We return to this in section 6 of this paper.

Rolls (1998) presents a different perspective on the psychology of the emotions. Brains are designed around reward and punishment (reinforcers) evaluation systems. Rather than reinforcing particular behavioural patterns of responses (behaviourism), the reinforcement mechanisms work in terms of cognitive activity such as goals and motivation. Emotions are states elicited by reinforcers. These states are more encompassing than those states associated with feelings of emotion. Emotions have many functions including the priming of reflexive behaviors associated with the autonomic and endocrine system, the establishment of motivational states, the facilitation of memory processing (storage and control) and maintenance of the "*persistent and continuing motivation and direction of behavior*". In effect Rolls suggests that the neuropsychological evidence supports the conjecture that emotions provide the glue that binds mind and personality.

Psychiatry, philosophy and emotion

Wollheim (1999) distinguishes two aspects of mental life in his analysis of the emotion: the phenomena of mental states and mental dispositions. Three very general properties characterise mental phenomena: intentionality, subjectivity and grades of consciousness (conscious, preconscious and unconscious).

Aaron and others?

Theoretical Framework

The theory presented here is rudimentary and sketchy. It builds on aspects of the work of presented above, and emphasises the interplay of cognition and emotion through motivation and peturbance. Earlier we used a psychological definition of emotion that referred to both cognitive (appraisal) and physiological factors. Emotions, in socio-biological agents, are affective mental (appraisal) states and processes, and any computational model of emotion must attempt to meet similar specifications. In moving towards a model of emotion that will be computationally tractable, we would like to minimise the extent of the model (ontological parsimony), and yet still remain somewhat true to the underlying psychological and physiological evidence. Sadness and happiness are antipathetic, being reflections of each other, or extremes on one dimension. Here we shall use the term sobriety, with sadness and happiness either side of a neutral state. The resulting four dimensional model is computationally tractable, and maps onto our ideas for architectures for minds. A further salient feature of these definitions of emotion is that they are described in terms of goals and roles. This enables emotions to be defined over different levels of an architecture for mind using different categories of behaviors. Furthermore, if emergent behaviors (that are related to emotions) are to be recognised and

managed then we can ensure that there is design synergy across the different layers of the architecture.

Emotions can be unconscious and managed by the autonomic nervous system and its biological substrate (including the endocrine systems). Emotions can move into the conscious mind or be invoked at that level (through cognitive appraisal of agent, object or event related scenarios). Emotions can be instantiated by events both internal and external at a number of levels of abstraction, whether primary (genetic and/or ecological drives) or by events that require substantive cognitive processing. Goleman (1995) discusses emotional high-jacking at length. An analogy from visual perception is the autonomous reflex exhibited by a frog in response to small black objects moving across its visual field. Emotions are temporally short, although emotional states resulting from successive waves of emotions can be phenomenologically more enduring. Emotions can be casually inter-related and cause other events. Drives and motivations are highly interlinked with emotions. These can embody some representation (not necessarily semantic) and in effect relate short-term emotive states to temporally global processes. For example,

Experimental computational work

The architecture for computational mind that we are developing (Davis 1996; 2000) is based on ideas developed within the Cognition and Affect group at Birmingham (Beaudoin and Sloman 1993;Sloman 1994), although perhaps differing from the latest thoughts from that group (Sloman 1999).

Future work

A better understanding of the relations between emotion, cognition, mind and consciousness.

Better HCI?

Intuitive reasoning?

Affective computation?

References

Beaudoin, L.P. and A. Sloman, A study of motive processing and attention, In: *Prospects for Artificial Intelligence,* Sloman, Hogg, Humphreys, Partridge and Ramsay (eds), IOS Press, 1993.

Davis, D.N., Reactive and motivational agents. In: *Intelligent Agents III*, J.P. Muller, M.J. Wooldridge & N.R. Jennings (Eds.), Springer-Verlag, 1996.

Davis, D.N., T. Chalabi and B. Berbank-Green, Towards an architecture for artificial life agents: II, In: M. Mohammadian (Editor), *New Frontiers in Computational Intelligence and Its Applications*, ISO Press, 1999.

Davis, D.N., Modelling emotion in computational agents, *Paper submitted to ECAI2000*, 2000.

Ekman, P., and R.J. Davidson (Eds.), *The Nature of Emotion*, Oxford University Press, 1994.

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

Gilbert, N. and Conte, R., *Artificial Societies: The computer simulation of social life*, UCL Press, (1995).

Goleman, D.P, *Emotional Intelligence*, Bloomsbury Publishing, 1995.

Gunderson, K., Mentality and Machines (2e), Croom Helm, 1985.

Harré, R., Emotion and memory: the second cognitive revolution. In: *Philosophy, Psychology and Psychiatry*, A.P. Griffiths (ed.), Cambridge University Press, (1994).

Hegselmann R. and Flache, A., Understanding complex social dynamics: A plea for cellular automata based modelling. *Journal of Artificial Societies and Social Simulation*, Vol. 1, No3, (1998).

Lazurus, R.S., Emotion and Adaptation, Oxford University Press, 1991.

Lindsay, P.H. and D.A. Norman, *Human Information Processing: An Introduction to Psychology*, Academic Press, 1972.

McCarthy J., Making robots conscious of their mental states. *Machine Intelligence 15*, Oxford, 1995.

Newell A., *Unified Theories of Cognition*, Harvard University Press, 1990.

Oatley, K. and Jenkins, J.M, *Understanding Emotions*, Blackwell, 1996.

Ortony, A., G.L. Clore and A. Collins, *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

Picard, R. *Affective Computing*, MIT Press, 1997.

Port R.F. and T. Van Gelder (Editors), *Mind As Motion*, MIT Press, 1995.

Power, M. and T. Dalgleish, *Cognition and Emotion:From Order to Disorder*, LEA Press, 1997.

Rolls, E.T., *The Brain and Emotion*, Oxford University Press, 1998.

Sloman, A. and M. Croucher, Why robots will have emotions. *Proceedings of IJCAI7*, 197-202, 1987.

Sloman, A. Explorations in design space. In *Proceedings 11th European Conference on AI*, Amsterdam, 1994.

Sloman, A. Architectural requirements for human-like agents both natural and artificial, In *Human Cognition and Social Agent Technology*, K. Dautenhahn, Benjamins Publishing, 1999.

Wollheim, R., *On The Emotions*, Yale University Press, 1999.