FA3

International Journal of Machine Consciousness
Vol. 1, No. 2 (2009) 203−211
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

EMOTIONAL COGNITIVE STEPS TOWARDS CONSCIOUSNESS

WILL N. BROWNE* and RICHARD J. HUSSEY†

Cybernetics, University of Reading, Reading, Berkshire, UK
**w.n.browne@reading.ac.uk*
†*r.j.hussey@reading.ac.uk*

The academic journey to a widely acknowledged Machine Consciousness is anticipated to be an emotional one. Both in terms of the active debate provoked by the subject and a hypothesized need to encapsulate an analogue of emotions in an artificial system in order to progress towards machine consciousness. This paper considers the inspiration that the concepts related to emotion may contribute to cognitive systems when approaching conscious-like behavior. Specifically, emotions can set goals including balancing explore versus exploit, facilitate action in unknown domains and modify existing behaviors, which are explored in cognitive robotics experiments.

Keywords: Artificial emotions; learning classifier systems; cognitive robotics; affective computing; analogues of emotion.

1. Introduction

It may be argued that evolution has led to consciousness in human beings. However, this process has taken millennia, which is not desirable when attempting to create artificial conscious-like behaviors. Thus, rather than start with a tabula rasa, it may be preferable to start with analogues of evolved human traits in an attempt to speed-up the creation of useful behaviors, e.g., emotions [Fellous and Arbib, 2005; Vallverdú and Casacuberta, 2009].

Emotions are developmental, i.e., only some emotions are present at birth, by nine months all basic emotions are present, self-awareness emotions (e.g., embarrassment), 18−24 months, and evaluated emotions (guilt) develop by 2−3 years. This trade-off between nature and nurture of artificial emotions for cognitive (leading to conscious) robotics invites exploration with initial investigations outlined in this paper. Firstly, is the nature of artificial emotions useful to a robot, such that it provides functionality not easily obtained by other means? Secondly, can emotions be tuned by nurture through interaction with a given environment in order to improve their usefulness to an agent?

A broad overview of functionality is "Emotions are reflections of the adaptations that animals make to universal problems" [Plutchik, 1991], where the universal problems of adaptations are temporality, identity, hierarchy and territoriality. Feedback from the environment [Breazeal, 2004] is used for both communication [Breazeal, 2004; Michaud *et al.*, 2001] and control [Arbib and Fellous, 2004; Di Paolo and Iizuka, 2004; Scheutz, 2004; Takeno *et al.*, 2005]. Fellous concludes that it may be "... more fruitful to focus on function of emotions not what they are", which is the focus adopted by this work [Fellous, 2004].

2. Background to Emotions

Theories of emotional functionality are over a hundred years old, including Darwin [1871] and James [1884]. More recently, Aleksander and Morton's [2007] emotional architecture model shows both afference and efference. Schachter's cognitive theory model shows efference then afference [Scherer, 1988]. There is an important difference between afference feedback and efference signals. Re-experiencing the emotional context of a state can affect the decision taken, so is a form of emotional feedback (see somatic marker hypothesis [Damasio, 1996]). Similarly, being in an emotional state biases the decision-making with this signal, then affects the output [Rolls, 1999].

The purpose of emotion is also debated, with a general categorization provided by [Michaud *et al.*, 2001]:

- to adapt to limitations
- for managing social behavior
- for interpersonal communication

This is not a definitive list as emotion has also been linked to the memory of facts, which is improved when the facts are learnt in connection with an emotion (to a limit). Also, there is a strong link between emotion and decision-making and other frontal lobe cognitive functions, e.g., working memory [Bechara *et al.*, 2000]. Many alternative viewpoints exist, e.g., [Rolls, 1999; Cahill *et al.*, 1995; Hamann, 2001].

Similar to the differing viewpoints on emotional functionality there are varying types, descriptions and definitions of emotions themselves. A well known classification is by Plutchik [1991] whose classification is based on purpose:

Temporality: *joy/sadness*
Identity: *acceptance/rejection*
Hierarchy: *anger/fear*
Territoriality: *expectation/surprise*

Anthropology studies by Ekman [1999] have identified classes of emotion with similarities across cultures. The 1972 list of anger, sadness, happiness, fear, disgust, surprise was expanded with contempt and embarrassment and then rewritten with 15 base emotions in 1999.

2.1. *Existing emotional models*

Computational models of emotion exist for both neuro-scientific understanding and for cognitive robotic control [Kawamura and Browne, 2009] (for standard architectures see [Kieras and Meyer, 1997; Laird *et al.*, 1987]). In the former class includes emotional learning in the amygdala [Moren and Balkenius, 2000], which is based on the Amygdalo-orbitofrontal system expounded by Rolls [1999] and LeDoux [1996]. The balance of inhibitory with excitory signals is important in the simulations [Shanahan, 2006]. John Taylor's group models the interaction of attention and emotion, including the enhancement of perception caused by emotional cues [Fragopanagos *et al.*, 2006].

A bridge between these types is the models of Fellous whose behavioral investigation contains an organization where the potential for emotional control (through neuro-modulation) increases with higher-level cognition (from reflexes to drives to instincts to cognitions). Other advice for emotional modeling in cognitive robotics includes Clark and Grush's promotion of a minimal yet robust internal representation [Clark and Grush, 1999] and affective architectures [Sloman and Chrisley, 2005; Sloman and Logan, 1998].

Kawamura *et al.* [2006] develops the role of episodic memory and emotion for the cognitive robot ISAC where the emotion component is based on Haikonen's System Reactions Theory of Emotion (SRTE) [Haikonen, 2003]. The relation between emotions and system reaction is predefined, e.g., pain due to an external agent will result in an aggressive response. The DARE architecture [Macas *et al.*, 2001] again works with the double and parallel stimuli processing concept of LeDoux plus the somatic marker concept of Damasio. Campagne and Cardon [2003] approach an emotion model from similar inspiration using a multi-agent perspective in a massive simulation only.

3. Emotional Inspiration

If robots are to benefit from mechanisms that have a similar role to emotions it is suggested to use internal variables [Michaud *et al.*, 2001]. However, Fellous [2004] warns that an isolated emotion is simply an engineering hack, i.e., simply describing a single, isolated internal variable as an emotion could be descriptive or anthropomorphic, but not biologically inspired. Instead, inter-related emotions, expressed due to resource mobilization with context-dependent computations dependent on perceived expression is more realistic.

Thus robot-emotions should be built from the following guidelines [Fellous, 2004]:

- emotions are not a separate center that computes a value on some predefined dimension;
- emotions should not be a result of cognitive evaluation (if state then this emotion);
- emotions are not combinations of some prespecified basic emotion (emotions are not independent from each other);

- emotions should have temporal dynamics and interact with each other;
- system wide control of some of the parameters (of the many ongoing, parallel processes) that determine the robot behavior.

3.1. *Multidimensional and multimodal*

There are strong arguments that single production rules cannot work as a basis for emotional cognitive control, but to dismiss all "if … then …" symbolic systems based on these arguments appears premature.

To explore some of the arguments against production rules, consider efferent emotions where state s_1 evokes emotion e. This link may not be derived just from state to action (a) to reward (r) to emotion s_1-a_1-r_1-e as the episode may have been s_1-a_1- s_2-a_2-s_3-a_3-s_4-a_4-r-e or alternative complex sequences. Similarly, it cannot now be stated that "if s_1 then e" as when the next instance of s_1 is presented the system reasons on afferent as well as efference signals s_1 and e, which could produce a completely different action-reward-emotion sequence, e.g., if a need has been satisfied.

From Sec. 2 the types of emotions are limited. Each emotion may be considered as a dimension, which consists of a series of limit cycles from one extreme reflecting aversive reward to the other representing appetitive reward. "Negative" emotions are just as essential for the survival of the agent as "positive" emotions as they help to avoid aversive states. It is to be decided how these emotional values are changed, but it is considered that states, actions, rewards and even emotions themselves may influence the emotional level. Although it is wrong to have an isolated production rule that always says s_1-e, it is acceptable to have a production rule that says s_1 and e evokes e' and s_1-Δe, where the evoking and changing of e is not unique to s_1.

3.2. *Affectors and effectors*

A production rule model for emotions could be s and e-a evokes e and Δe. This can be complemented by direct rules, e.g., s_1-a or s_1-e and can be inhibited by other rules where these are hardwired or learnt.

Haikonen [2003] identifies five stages of consciousness with Step 3 relevant here. Emotions are considered to be for attention control, motivation; a shortcut template for style of action and affect learning. Thus selected states invoke emotions and the totality of emotions corresponds to rewards.

Forward modeling is necessary for cognition with efference copies being required to predict the world. It is considered that emotions are likely to be evoked when the model and world do not match. It may be considered that the world is its own best model [Brooks, 1985], but there is still a requirement for an internal model on how best to use the external model with sensors [Holland, 2003]. The evolutionary computation-based production rule system of Learning Classifier Systems (LCS) was originally proposed by Holland as a cognitive system [Holland, 1975]. Decades of research have enabled LCS to become an effective machine learning technique [Browne and Tingley, 2006; Lanzi, 2002; Butz, 2004].

3.3. *Generalization, abstraction and anticipations*

Learning requires memorization of perceptions from the environment in order to store useful behaviors.

(i) Irrelevant information from a perceived state must be removed. Attention must be focused on the important components of the state, which may be accomplished through learning generalizations.
(ii) Higher-order patterns must be abstracted from learnt episodes so that rules may be applied to similar situations.
(iii) Anticipatory models must then be built up linking future states to existing states with plausible actions.

LCS generalize by denoting irrelevant conditions in a state as "don't care" or by removing them from the production rule itself. Initial work has shown abstraction is possible and beneficial in appropriate environments [Browne *et al.*, 2008], but further work is required. This is also true for the determination of affordances, which is assisted by the matching of rules to states by LCS including forward planning. Anticipatory LCS implementations include ACS, ACSII and AgentP where *s-a-s* rules are autonomously formed [Zatuchna, 2005].

3.4. *Memory*

Many types of memory have been classified in humans, including short-term, long-term and working memory [Baxter and Browne, 2008]. However, much debate exists regarding the underlying biological mechanisms for the observed functional differences [Phillips and Noelle, 2005].

The ability to generalize and abstract does reduce the required memory, say compared with a Q-learning state table. However, to accurately model a practical world would still require much memory of rules leading to potentially slow searching, accessing and maintenance. Efficient matching by removing irrelevant conditions also assists, whilst improved anticipations could enable pre-searching of likely rules that might be activated in the near future. Importantly, the ability to associate emotions with rules could greatly reduce the search space, e.g., a sad agent may ignore many rules associated with joy.

Communication through modeling emotional states facilitates the ability to place other agents in an "out there" world. By mirroring these rules the agent can then place itself in the "out there" world too. These abilities form part of the arguments by Holland [2003] for building a consciousness [Browne and Tingley, 2006].

4. Results from Cognitive Robotics Experiments

Nature: Initially, a static rule-base was created manually including both afference and efference rules but no direct state-action links, see Fig. 1 [Browne and

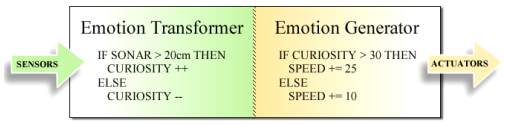

Fig. 1. An example of an emotion transformer and associated transformer.

Tingley, 2006]. Multiple states could affect an emotion and each of the five modeled emotions could effect an action.

Superimposed runs of an explore task are shown in Fig. 2. A non-emotional benchmark architecture produced almost identical paths when starting from the same position, which is common in deterministic systems. Introducing randomness to the controller was ineffective as the robot made little forward progress. However, when considering the paths generated by the emotional system, both runs were completely different, but equally effective — ∼90% exploration compared with ∼70% for the benchmark system.

Nurture: The emotion analogues selected for this problem domain were: Happiness (P+), Sadness (P−), Curiosity (I+), Anger (I−), Hope (D+), and Fear (D−). Where each emotional signal is related to satisfying the need by a proportional (P), integral (I) or derivative (D) relationship and may be appetitive (+) or aversive (−). An LCS is used to learn the effective use of emotions, i.e., best action for a given emotional state. Again the task was to explore, initially in a simple domain. Experiments showed that as the training progressed the area explored in a given time interval increased, see Fig. 3.

Examining the rules produced showed plausible learning had occurred. The autonomously identified fittest rule learnt:

$$\begin{aligned} 111 - \text{If } &((18 \Leftarrow \text{Happiness} \Leftarrow 88) \text{ and } (6 \Leftarrow \text{Sadness} \Leftarrow 23) \text{ and} \\ &(12 \Leftarrow \text{Curiosity} \Leftarrow 81) \text{ and } (17 \Leftarrow \text{Anger} \Leftarrow 85) \text{ and} \\ &(17 \Leftarrow \text{Hope} \Leftarrow 85) \text{ and } (8 \Leftarrow \text{Fear} \Leftarrow 52)) \ \{\text{Action} = 1\} \end{aligned}$$

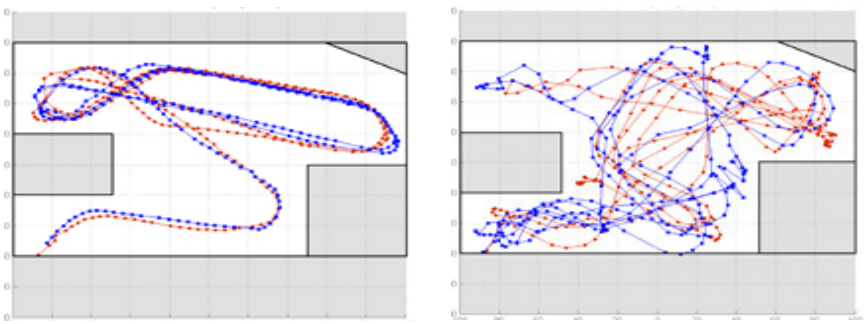

Fig. 2. A non-emotional agent architecture exploring a complex domain for three minutes (left). An emotional-based agent architecture exploring a complex domain for three minutes. Note the non-deterministic nature of the two runs (right).

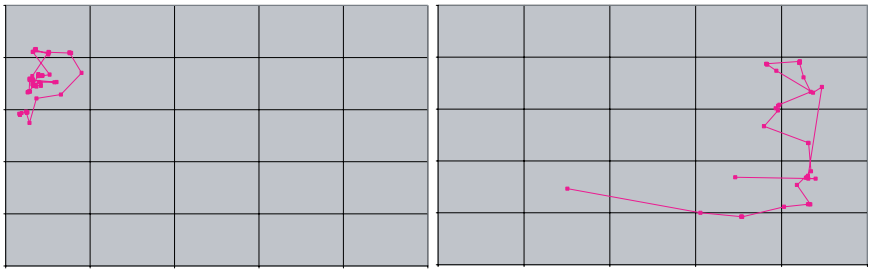

Fig. 3. Plot of the map data from LCS iteration 4 (left) and map data from LCS iteration 9 (right).

This corresponds to the behavior: agent is in the open space, then go forward slowly. As the motor actions were set arbitrarily, this suggests that the top range of speeds were too fast for the domain leading to collisions with the walls.

Due to the accuracy-based nature of the LCS used, it also formed aversive rules. The behavior to be most avoided is: if agent is trapped, then turn left fast (note that trapped corresponds to low values from the ultrasonic sensors).

5. Discussion

Production rule-based systems have potential for both affective and effective learning, especially when internal reward is linked to satisfying identified needs. The ability to reason on emotional states, including combinations of multiple emotions, is essential for generating realistic behaviors. Furthermore, the ability to set emotions, change emotions and form chains of rules adds to the temporal, including episodic, learning capabilities.

It was noted that the control actions of the robot-emotions were similar to the actions of conventional controllers. A simple analogy is that some emotions reacted proportionally to the input signal, others built up over time (integrated) and others reacted to the rate of change of a signal (differential). Proportional, integral and derivative (PID) is a standard industrial control strategy. It is worth considering the links to other conventional control strategies, such as filter-based techniques (e.g., lead control), model-based control and adaptive control to determine if natural emotions have similar actions that could be replicated artificially.

6. Conclusion

Inspiration from concepts related to emotion is likely to significantly contribute to cognitive systems when approaching conscious-like behavior. Specifically, emotions can set goals including balancing explore versus exploit, facilitate action in unknown domains and modify existing behaviors, which can be shown in initially simple cognitive robotics experiments. A need for predictive certainty coupled with emotional communication is postulated to lead to an agent approaching conscious behaviors when placing itself in an "out there" world.

References

Arbib, M. A. and Fellous, J.-M. [2004] "Emotions: From brain to robot," *Trends in Cognitive Sciences* **8**, 554–561.

Aleksander, I. and Morton, H. [2007] "Why axiomatic models of conscious?" *Journal of Consciousness Studies* **14**(7), 15–27.

Baxter, P. and Browne, W. [2008] "Towards a developmental memory-based and embodied cognitive architecture," *Epigenetic Robotics* **8**, University of Sussex, pp. 137–138.

Bechara, A., Damasio, H. and Damasio, A. [2000] "Emotion, decision-making and the orbitofrontal cortex," *Cerebral Cortex* **10**, 295–307.

Breazeal, C. [2004] "Function meets style: Insights from emotion theory applied to HRI," *IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews* **34**(2).

Brooks, R. A. [1985] "A robust layered control system for a mobile robot," Technical Report, Massachusetts Institute of Technology, Cambridge, MA.

Browne, W. N. and Tingley, C. [2006] "Developing an emotion-based architecture for autonomous agents," in *Third International Conference on Autonomous Robots and Agents* (*ICARA'06*), Palmerston North, New Zealand, pp. 225–230.

Browne, W. N., Scott, D. and Ioannides, C. [2008] "Abstraction for genetics-based reinforcement learning," in *Reinforcement Learning: Theory and Applications*, eds. Weber, C., Elshaw, M. and Mayer, N. M. (Advanced Robotic Systems Publishing, Vienna, Austria).

Butz, M. [2004] "Rule-base evolutionary online learning systems: Learning bounds, classification and prediction," PhD thesis, University of Illinois, Illinois.

Cahill, L., Babinsky, R., Markowitsch, H. J. and McGaugh, J. L. [1995] "The amygdala and emotional memory," *Nature* **377**(6547), 295–296.

Campagne, J. C. and Cardon, A. [2003] "Artificial emotions for robots using massive multiagent systems," *SID'03*, London.

Clark, A. and Grush, R. [1999] "Towards cognitive robotics," *Adaptive Behavior* **7**(1), 5–16.

Darwin, C. [1871] *The Descent of Man and Selection in Relation to Sex* (John Murray, London), reprinted in 1981 by Princeton University Press.

Damasio, A. [1996] "The somatic marker hypothesis and the possible functions of the prefrontal cortex," *Philosophical Transactions of the Royal Society B* **351**, 1413–1420.

Di Paolo, E. and Iizuka, H. [2008] "How (not) to model autonomous behavior," *BioSystems* **91**, 409–423.

Ekman, P. [1999] "Basic emotions," in *Handbook of Cognition and Emotion*, eds. Dalgleish, T. and Power, T. (John Wiley & Sons, Sussex), pp. 45–60.

Fellous, J.-M. [2004] "From human emotions to robotic emotions," American Association for Artificial Intelligence — Spring Symposium 3/2004, Stanford University.

Fellous, J.-M. and Arbib, M. A. [2005] *Who Needs Emotions? The Brain Meets the Robot* (Oxford University Press, USA).

Fragopanagos, N., Korsten, N. and Taylor, J. G. [2006] "A neural model of the enhancement of perception caused by emotional cues," in *Proceedings*.

Hamann, S. [2001] "Cognitive and neural mechanisms of emotional memory," *Trends in Cognitive Sciences* **5**, 394–400.

Haikonen, P. O. [2003] *The Cognitive Approach to Conscious Machines* (Imprint Academic, UK).

Holland, J. H. [1975] *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, MI).

Holland, O. (ed.) [2003] *Machine Consciousness* (Imprint Academic, UK).

James, W. [1884] "What is an emotion?" *Mind* **9**, 188–205.

Kawamura, K. *et al.* [2006] "From intelligent control to cognitive control," *11th International Symposium on Robotics and Applications (ISORA)*, Budapest, Hungary.

Kawamura, K. and Browne, W. N. [2009] "Cognitive robotics," in *Encyclopedia of Complexity and System Science*, ed. Meyers, B. (Springer, USA).

Kieras, D. E. and Meyer, D. E. [1997] "An overview of the EPIC architecture for cognition and performance with application to human−computer interaction," in *Human–Computer Interaction*.

Laird, J., Newell, A. and Rosenbloom, P. [1987] "Soar — An architecture for general intelligence, *Artificial Intelligence* **33**, 1−64.

Lanzi, P.-L. [2002] "Learning classifier systems from a reinforcement learning perspective," *Soft Computing: A Fusion of Foundations, Methodologies and Applications* **6**, 162−170.

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

Macas, M., Ventura, R., Custodio, L. and Pinto-Ferreira, C. [2001] "Experiments with an emotion-based agent using the DARE architecture," in *Proceedings of the Symposium on Emotion, Cognition, and Affective Computing, AISB'01 Convention*.

Michaud, F., Robichaud, E. and Audet, J. [2001] "Using motives and artificial emotions for long-term activity of an autonomous robot," in *Proceedings of the 5th International Conference on Autonomous Agents*, Montreal, Quebec, Canada, pp. 188−189.

Moren, J. and Balkenius, C. [2000] "A computational model of emotional learning, in the amygdala," in *Proceedings of the 6th International Conference on the Simulation of Adaptive Behavior*, Cambridge.

Phillips, J. L. and Noelle, D. C. [2005] "A biologically inspired working memory framework for robots," in *IEEE International Workshop on Robots and Human Interactive Communication*, Nashville, TN, pp. 599−604.

Plutchik, R. [1991] *The Emotions* (University Press of America, Lanham, MD).

Rolls, E. T. [1999] *The Brain and Emotion* (Oxford University Press, USA).

Scherer, K. R. [1988] "Criteria for emotion-antecedent appraisal: A review," in *Cognitive Perspective on Emotion and Motivation*, eds. Hamilton, V., Bower, G. and Frijda, N. (Kluwer Academic), pp. 89−126.

Scheutz, M. [2004] "Useful roles of emotions in artificial agents: A case study from artificial life," in *American Association for Artificial Intelligence AAAI'04*, pp. 42−48.

Shanahan, M. P. [2006] "A cognitive architecture that combines internal simulation with a global workspace," *Consciousness and Cognition* **15**, 433−449.

Sloman, A. and Logan, B. [1998] "Cognition and affect: Architectures and tools," in *Proceedings of the 2nd International Conference on Autonomous Agents*, New York, pp. 471−472.

Sloman, A. and Chrisley, R. [2005] "More things than are dreamt of in your biology: Information processing in biologically-inspired robots," *Cognitive Systems Research* **6**.

Takeno, J. *et al.* [2005] "Experiments and examination of mirror image cognition using a small robot," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Espo, Finland, pp. 493−498.

Vallverdú, J. and Casacuberta, D. [2009] *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence* (Hershey, New York).

Zatuchna, Z. V. [2005] "Agent P: A learning classifier system with associative perception in maze environments," Thesis, School of Computing Sciences, University of East Anglia.

FA3