Architectures for Modeling Emotions. AAAI Spring Symposia Technical Series Technical Reports SS-04-02

Feelings and Emotions as Motivators and Learning Facilitators

Stan Franklin and Lee McCauley

Computer Science Division
The University of Memphis
Memphis, TN 38152, USA
franklin@memphis.edu, t-mccauley@memphis.edu

Abstract

The use of feelings and emotions as primary motivators and facilitators of several types of learning within the IDA architecture is described.

Introduction

Feelings in humans include hunger, thirst, pain, tiredness, depression, etc. Emotions, such as fear, anger, joy, sadness, etc., are feelings with cognitive content (Johnston 1999). Feelings, including emotions, are nature's means of implementing motivations for actions in humans and other animals. They have evolved so as to adapt us to regularities in our environments.

Artificial feelings and emotions are beginning to play an increasingly important role as mechanisms for primary motivations in software agents and robots, as well as facilitators of learning in these systems (Marsella & Gratch 2002, Langley et al. 2003). Here we present a case study of such feelings and emotions playing both roles in an intelligent software agent performing a practical, real world task. In this agent they are actively involved in every instance of action selection, and at least potentially involved in each learning event. The pervasive, central role that feelings and emotions play in the control structure of this software agent mimics the roles they play in human cognition, and gives rise to clarifying hypotheses about human decision making and several forms of human learning.

The IDA Model

IDA provides a conceptual (and computational) model of cognition (Franklin 2000, 2001b) implemented as a software agent (Franklin & Graesser 1997). IDA "lives" on a computer system with connections to the Internet and various databases, and does personnel work for the US Navy, performing all the specific personnel tasks of a human (Franklin 2001a). In particular, IDA negotiates with sailors in natural language, deliberates, and makes voluntary action selections in the process of finding new jobs for sailors at the end of their current tour of duty. IDA completely automates the work of Navy personnel agents (detailers).

The IDA model implements and fleshes out Global Workspace theory (Baars 1988, 2002), which suggests points when beginning a new paragraph, unless the paragraph that conscious events involve widespread distribution of focal information needed to recruit neuronal resources for problem solving. The IDA implementation of GW theory yields a fine-grained functional account of the steps involved in perception, several kinds of memory, consciousness, context setting, and action selection. Cognitive processing in IDA consists of continually repeated traversals through the steps of a *cognitive cycle* (Baars & Franklin 2003, Franklin et al. in review), as described below.

The IDA architecture includes modules for perception (Zhang, et al. 1998), various types of memory (Anwar and Franklin. 2003, Franklin et al. in review), "consciousness" (Bogner, Ramamurthy and Franklin. 2000), action selection (Negatu and Franklin. 2002), constraint satisfaction (Kelemen, Liang, and Franklin. 2002), deliberation (Franklin 2000a), and volition (Franklin 2000a). The mechanisms of these modules are derived from several different "new AI" sources (Hofstadter and Mitchell. 1994, Jackson 1987, Maes 1989).

IDA senses strings of characters from email messages and databases, and negotiates with sailors via email. The computational IDA is a running software agent that has been tested and demonstrated to the satisfaction of the Navy. Detailers observing the testing commented that "IDA thinks like I do."

In addition to the computational model, we will also speak of the conceptual IDA model, which includes additional capabilities that have been designed but not implemented, including mechanisms for feelings and emotions.

The IDA conceptual model contains several different memory systems. Perceptual memory enables identification, recognition and categorization, including of feelings. Working memory provides preconscious buffers as a workspace for internal activities. Transient episodic memory is a content-addressable associative memory with a moderately fast decay rate. It is to be distinguished from autobiographical memory, a part of long-term associative memory. Procedural memory is long-term memory for skills.

Much of the activity within IDA is accomplished by codelets, small pieces of code that each performs one specialized, simple task. Codelets often play the role of demons waiting for a particular type of situation to occur and then acting as per their specialization. Codelets in the IDA model implement the processors postulated by global workspace theory. Neurally they can be thought of as cell assemblies or neuronal groups (Edelman1987, Edelman and Tononi 2000). The various sorts of codelets, perceptual, attentional, behavioral, expectational, will be described below.

The Cognitive Cycle

The IDA model suggests a number of more specialized roles for feelings in cognition, all combining to produce motivations and to facilitate learning. Here we describe IDA's cognitive cycle in nine steps, emphasizing the roles played by feelings and emotions:

1. ***Perception***. Sensory stimuli, external or internal, are received and interpreted by perception creating meaning. Note that this stage is unconscious.
    a. Early perception: Input arrives through senses. Specialized perception codelets descend on the input. Those that find features relevant to their specialty activate appropriate nodes in the slipnet (a semantic net with activation).
    b. Chunk perception: Activation passes from node to node in the slipnet. The slipnet stabilizes, bringing about the convergence of streams from different senses and chunking bits of meaning into larger chunks. These larger chunks, represented by meaning nodes in the slipnet, constitute the percept. *Pertinent feeling/emotions are identified (recognized) along with objects and their relations by the perceptual memory system. This could entail simple reactive feelings based on a single input or more complex feelings requiring the convergence of several different percepts.*
2. ***Percept to Preconscious Buffer***. The percept, including some of the data plus the meaning, is stored in preconscious buffers of IDA's working memory. These buffers may involve visuo-spatial , phonological, and other kinds of information. *Feelings/emotions are part of the preconscious percept written during each cognitive cycle into the preconscious working memory buffers.*
3. ***Local Associations***. Using the incoming percept and the residual contents of the preconscious buffers as cues including emotional content, local associations are automatically retrieved from transient episodic memory and from long-term associative memory. The contents of the preconscious buffers together with the retrieved local associations from transient episodic memory and long term associative memory, roughly correspond to Ericsson and Kintsch's long-term working memory (1995) and Baddeley's episodic buffer (2000). *Feelings/emotions are part of the cue that results in local associations from transient episodic and declarative memory. These local associations contain records of the agent's past feelings/emotions in associated situations*.
4. ***Competition for Consciousness***. Attention codelets, whose job it is to bring relevant, urgent, or insistent events to consciousness, view long-term working memory. Some of them gather information, form coalitions and actively compete for access to consciousness. The competition may also include attention codelets from a recent previous cycle. *Present and past feelings/emotions influence the competition for consciousness in each cognitive cycle. Strong affective content strengthens a coalition's chances of coming to consciousness.*
5. ***Conscious Broadcast***. A coalition of codelets, typically an attention codelet and its covey of related information codelets carrying content, gains access to the global workspace and has its contents broadcast. This broadcast is hypothesized to correspond to phenomenal consciousness. *The conscious broadcast contains the entire content of consciousness including the affective portions.* The contents of perceptual memory are updated in light of the current contents of consciousness, *including feelings/emotions*, as well as objects, and relations. *The stronger the affect, the stronger the encoding in memory.* Transient episodic memory is updated with the current contents of consciousness, *including feelings/emotions*, as events. *The stronger the affect, the stronger the encoding in memory.* (At recurring times not part of a cognitive cycle, the contents of transient episodic memory are consolidated into long-term declarative memory.) Procedural memory (recent actions) is updated (reinforced) with the strength of the reinforcement influenced by the strength of the affect.
6. ***Recruitment of Resources***. Relevant behavior codelets respond to the conscious broadcast. These are typically codelets whose variables can be bound from information in the conscious broadcast. If the successful attention codelet was an expectation codelet calling attention to an unexpected result from a previous action, the responding codelets may be those that can help to rectify the unexpected situation. Thus consciousness solves the relevancy problem in recruiting resources. *The affective content (feelings/emotions) together with the cognitive content help to attract relevant resources*

(processors, neural assemblies) with which to deal with the current situation.

7. ***Setting Goal Context Hierarchy***. The recruited processors use the contents of consciousness, *including feelings/emotions*, to instantiate new goal context hierarchies, bind their variables, and increase their activation. It's here that feelings/emotions directly affect motivation. They determine which terminal goal contexts receive activation and how much. *It is here that feelings and emotions most directly implement motivations by helping to instantiate and activate goal contexts.* Other, environmental, conditions determine which of the earlier goal contexts receive additional activation.
8. ***Action Chosen***. The behavior net chooses a single behavior (goal context), perhaps from a just instantiated behavior stream or possibly from a previously active stream. *This selection is heavily influenced by activation passed to various behaviors influenced by the various feelings/emotions*. The choice is also affected by the current situation, external and internal conditions, by the relationship between the behaviors, and by the residual activation values of various behaviors.
9. ***Action Taken***. The execution of a behavior (goal context) results in the behavior codelets performing their specialized tasks, which may have external or internal consequences. This is IDA taking an action. The acting codelets also include an expectation codelet (see Step 6) whose task it is to monitor the action and to try and bring to consciousness any failure in the expected results.

We suspect that cognitive cycles occur five to ten times a second in humans, overlapping so that some of the steps in adjacent cycles occur in parallel (Baars, B. J., and S. Franklin. 2003. How conscious experience and working memory interact. *Trends in Cognitive Science* 7:166-172.). Seriality is preserved in the conscious broadcasts.

Related Work

The IDA architecture differs significantly from such other cognitive architectures such as SOAR (Laird et al. 1987) and ACT-R (Anderson & Lebiere 1998) in that it is not a unified theory of cognition in the sense of Newell (1990). Rather, its various modules are implemented by a variety of different mechanisms including the copycat architecture, sparse distributed memory, pandemonium theory and behavior nets (Franklin 2001b). Though the IDA architecture contains no production rules and no neural networks, it does incorporate both symbolic and connectionist elements. The IDA architecture allows feelings and emotions to play a central role in perception, memory, "consciousness" and action selection as will be described below. IDA's "consciousness" mechanism, based on Global Workspace theory, resembles a blackboard system (Nii 1986), but there is much more to the IDA architecture having to do with the interaction of its various modules. This interaction is described in the cognitive cycle detailed below.

The IDA architecture can be viewed as a specification of the more general CogAff architecture of Sloman (Wright et al. 1996). It has reactive and deliberative mechanisms but, as yet, no meta-management. There is a superficial resemblance between IDA and the ACRES system (Moffat et al. 1993) in that both interact with users in natural language. They are also alike in using emotions to implement motivations. Rather than viewing emotions as implementing motivations for the selection of actions on the external environment, Marsella and Gratch study their role in internal coping behavior (Marsella & Gratch 2002). From our point of view this is a case of emotions implementing motivation for internal actions as also occurs in the IDA conceptual model. The ICARUS system also resembles a portion of the IDA conceptual model in that it uses affect in the process of reinforcement learning (Langley et al. 1991).

Conclusions

Being generated from order-of-magnitude one hundred thousand lines of code, IDA is an exceedingly complex software agent. Thus from the usefulness of artificial feelings and emotions in the IDA architecture one would not jump to the conclusion that they would play useful roles in a more typical order-of-magnitude simpler software agent or robotic control structure. Besides, feelings and emotions are, as yet, only part of the IDA conceptual model, and have not been implemented. Significant difficulties could conceivably occur during implementation. That artificial feelings and emotions seem to play significantly useful roles in the conceptual version of IDA's cognitive cycles is not a conclusive argument that they will do so in simpler, implemented artificial autonomous agents.

Still, the IDA model suggests that software agents and robots can be designed to use feelings/emotions to implement motivations, offering a range of flexible, adaptive possibilities not available to the usual more tightly structured motivational schemes such as causal implementation, or explicit drives and/or desires/intentions.

So, what can we conclude? Note that the computational IDA performs quite well with explicitly implemented drives rather than with feelings and emotions. It is possible that a still more complex artificial autonomous agent with a task requiring more sophisticated decision making would require them, but we doubt it. Explicit drives seem likely to suffice for quite flexible action selection in artificial agents, but not in modeling biological agents. It appears that feelings and emotions come into their own in agent architectures requiring sophisticated learning. This case study of the IDA architecture seems to suggest that artificial feelings and emotions can be expected to be of most use in software agents or robots in which online learning of facts and/or skills is of prime importance. If this requirement were present, it would make sense to also implement primary motivations by artificial feelings and emotions.

References

Anwar, A., and S. Franklin. 2003. Sparse Distributed Memory for "Conscious" Software Agents. *Cognitive Systems Research* 4:339-354.

Anderson, J. R., and C. Lebiere. 1998. *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. J. 2002. The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Science* 6:47–52.

Baars, B. J., and S. Franklin. 2003. How conscious experience and working memory interact. *Trends in Cognitive Science* 7:166–172.

Baddeley, A. D. 2000. The episodic buffer: a new component of working memory? *Trends in Cognitive Science* 4:417-423.

Bogner, M., U. Ramamurthy, and S. Franklin. 2000. Consciousness" and Conceptual Learning in a Socially Situated Agent. In Human Cognition and Social Agent Technology, ed. K. Dautenhahn. Amsterdam: John Benjamins.

Edelman, G. M. 1987. *Neural Darwinism*. New York: Basic Books.

Edelman, G. M., and G. Tononi. 2000. *A Universe of Consciousness*. New York: Basic Books.

Ericsson, K. A., and W. Kintsch. 1995. Long-term working memory. *Psychological Review* 102:211-245.

Franklin, S. 2000. Modeling Consciousness and Cognition in Software Agents. In *Proceedings of the Third International Conference on Cognitive Modeling, Groeningen, NL, March 2000*, ed. N. Taatgen. Veenendal, NL: Universal Press.

Franklin, S. 2000a. Deliberation and Voluntary Action in 'Conscious' Software Agents. *Neural Network World* 10:505-521.

Franklin, S. 2001a. Automating Human Information Agents. In *Practical Applications of Intelligent Agents*, ed. Z. Chen, and L. C. Jain. Berlin: Springer-Verlag.

Franklin, S. 2001b. Conscious Software: A Computational View of Mind. In *Soft Computing Agents: New Trends for Designing Autonomous Systems*, ed. V. Loia, and S. Sessa. Berlin: Springer (Physica-Verlag).

Franklin, S., and A. C. Graesser. 1997. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Intelligent Agents III*. Berlin: Springer Verlag.

Franklin, S., B. J. Baars, U. Ramamurthy, and M. Ventura. in review. The Role of Consciousness in Memory. .

Hofstadter, D. R., and M. Mitchell. 1994. The Copycat Project: A model of mental fluidity and analogy-making. In *Advances in connectionist and neural computation theory, Vol. 2: logical connections*, ed. K. J. Holyoak, and J. A. Barnden. Norwood N.J.: Ablex.

Jackson, J. V. 1987. Idea for a Mind. *Siggart* Newsletter, 181:23-26.

Johnston, V. S. 1999. *Why We Feel:The Science of Human Emotions*. Reading MA: Perseus Books.

Kelemen, A., Y. Liang, and S. Franklin. 2002. A Comparative Study of Different Machine Learning Approaches for Decision Making. In *Recent Advances in Simulation, Computational Methods and Soft Computing*, ed. E. Mastorakis. Piraeus, Greece: WSEAS Press.

Laird, E. J., A. Newell, and Rosenbloom P. S. 1987. SOAR: An Architecture for General Intelligence. *Artificial Intelligence* 33:1–64.

Langley, P., K. B. McKusick, J. A. Allen, W. F. Iba, and K. Thompson. 1991. A design for the ICARUS architecture. *ACM SIGART Bulletin* 2:104–109.

Langley, P., D. Shapiro, M. Aycinena, and M. Siliski. 2003. A value-driven architecture for intelligent behavior. In Proceedings of the IJCAI-2003 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions.

Maes, P. 1989. How to do the right thing. *Connection Science* 1:291-323.

Marsella, S., and J. Gratch; 2002. A Step Towards Irrationality: Using Emotion to Change Belief. *1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*. Bologna, Italy. July 2002.

Moffat, D., N. H. Frijda, and R. H. Phaf. 1993. Analysis of a computer model of emotions. In *Prospects for artificial intelligence*, ed. A. Sloman, D. Hogg, G. R. Humphreys A, and A. Ramsay. Amsterdam: IOS Press.

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

Negatu, A., and S. Franklin. 2002. An action selection mechanism for 'conscious' software agents. Cognitive Science Quarterly 2:363-386.

Nii, H. P. 1986. The Blackboard Model of Problem Solving and the Evolution of Blackboard Architectures. *The AI Magazine* Summer 1986:38–53.

Wright, I., A. Sloman, and L. Beaudoin. 1996. Towards a Design-Based Analysis of Emotional Episodes. *Philosophy Psychiatry and Psychology* 3:101–126.

Zhang, Z., S. Franklin, B. Olde, Y. Wan, and A. Graesser. 1998. Natural Language Sensing for Autonomous Agents. In *Proceedings of IEEE International Joint Symposia on Intellgence Systems 98*.