Project Joshua Blue: Design Considerations for Evolving an Emotional Mind in a Simulated Environment

Nancy Alvarado, Sam S. Adams, Steve Burbeck, Craig Latta

IBM, Thomas J. Watson Research Center

Abstract

This paper contrasts the implementation of motivation and emotion in Project Joshua Blue with current approaches such as Breazeal's (2001) sociable robots. Differences in our implementation support our different goals for model performance and are made possible by a novel system architecture.

Overview of Joshua Blue

Project Joshua Blue applies ideas from complexity theory and evolutionary computational design to the simulation of mind on a computer. The goal is to enhance artificial intelligence by evolving such capacities as common sense reasoning, natural language understanding, and emotional intelligence, acquired in the same manner as humans acquire them, through learning situated in a rich environment.

This project is in its beginning stages. A simple model of mind has been implemented in a limited virtual environment. Even in this first, simple model, emotion and motivation are not separate programs or subroutines but are integral to the basic functions of mind and have a constant and pervasive influence on all mental activity. We believe that the complex social behaviors observed in humans will emerge as capacities of mind from the exercise of emotion and motivation in social environments. More importantly, however, we believe that integrating emotion and motivation with cognition is essential to achieving common sense reasoning and natural language understanding, to autonomous learning, and to goal-setting. In short, this integration is essential to endowing a computer with the ability to comprehend "meaning" as humans do.

The main goal of Project Joshua Blue is to achieve cognitive flexibility that approaches human functioning. We believe emotion is a mediating mechanism that permits flexible assignment of meaning and significance in different contexts, coupled with a way of navigating a dynamic environment. To do this, emotion itself must not be fixed in its relationships, but free to associate variably with environmental stimuli and internal mental events.

That emotion guides cognition is contrary to the theory of emergent emotions, where emotion is in the eyes of the observer, attributed to a robot or other entity based on its interaction with the environment (Shibata, 1999). Further, it is contrary to the modularity proposed by Brooks (1986) and others. We believe isolated or limited implementations of emotional capacity must result in limited functionality.

Comparisons with Sociable Robots

Sociable robots are relevant to our project because we expect Joshua Blue to ultimately learn through embeddedness in a social environment. Breazeal's (2001) promising approach to implementing emotion in robots appears to directly instantiate emotion using logic. She gives sociable robots emotion by specifying: (1) the conditions under which certain affective states arise, (2) criteria for arbitrating among competing emotions, and (3) the instrumental and expressive behaviors resulting from each affect (Breazeal, 2001). In her model, the releasers for affect must be specified, which implies that the designer must anticipate the possible drives or goals and define emotion-evoking situations. The response to those situations is fixed once the emotion is identified, and there is no ability for the robot to override it. As yet, there is no reflexivity, self-awareness, consciousness of emotional state, subjective feeling beyond what is simulated behaviorally, and there is no ability to maintain affective privacy or engage in impression management by dissembling. Without ability to selectively inhibit behavior, there is no possibility of conforming to social display rules or using affective expression instrumentally through deceit.

This approach, and similar rule-based or logic-based implementations of emotional intelligence have accomplished an amazing amount of functionality. Their designers clearly intend to expand emotional competence, but in doing so they are likely to encounter the same resource limitations as are faced by those using rule-based approaches to reasoning or knowledge management (Brooks, 1986). Thus, for Joshua Blue we sought a different approach to implementing both emotion and other cognitive abilities, beyond rule-based approaches, neural nets that also have limitations, and statistical approaches to simulating cognition.

Joshua Blue incorporates an emotional model derived from current emotion theory, and is thus superficially similar to models implemented by Breazeal and others. Like such models, our system includes valence and arousal, homeostasis, and drive states, but it also includes proprioception and a pain/pleasure system. The system

architecture is based on a semantic network of nodes connected by wires along which activation spreads (Quillian, 1966; Collins & Loftus, 1975). In traditional spreading activation models, the length of wires captures semantic distance. In our model, the conductance of wires is adjusted dynamically based on the emotional context. This design permits cognitive processes and mental representations to be continuously influenced by affect. Further, like many current approaches, the system is motivated and guided by affect to navigate its environment and acquire meaning through principles of learning.

A key difference between our model and current approaches is that emotion is implemented in both global and specific ways. Like Breazeal's (2001) model, our system uses tags for valence, but not for arousal or stance. When a node is activated, its valence influences the valence of the entire system, but is also modified by the global affect of the system. This makes possible emotionally driven shifts in cognition. Arousal guides attention and determines the strength of associations formed, interacting with valence to tag specific objects with additional significance. Affective weighting is important in determining which associated objects will cross a threshold for consciousness or be retained in memory. Proprioceptors for affect were implemented to permit the system to introspect on its own global affective state, to be aware of the affect associated with a specific set of objects, and to experience pain and pleasure. This latter constitutes the reward and punishment system that guides exploratory behavior, generates expectations and ultimately motivates goal-directed behavior.

Unlike Breazeal, we have made no attempt to instantiate Ekman's basic emotions. We believe such states will emerge from learning and social interaction, providing a test of current emotion theories. Aside from the motive to seek pleasure and avoid pain, we are also incorporating a more complex structure of drives. We reserve the term "drive" for innate or hard-wired motives essential to survival, such as hunger or thirst in humans. Beyond that, the coupling of affect and experience should result in the formation of acquired motives and associated goals that have attained emotional significance through social learning (Reeve, 1997).

Breazeal (2001) uses positive emotions to signal that activity toward a goal can terminate and resources can be released. Neuropsychological evidence supports the idea that pleasure indefinitely sustains seeking or approach behavior, while other mechanisms indicate satiety (Panksepp, 1998). In our system, positive affect or pleasure arises not only from consummatory behavior but also from the exercise of certain intrinsic cognitive processes that require no homeostatic regulation (e.g., autonomy and control, familiarity and liking, competence and self esteem, social attachment). Pleasure is thus not a signal to terminate a drive state but a motive for approach behaviors. To terminate goals, our model incorporates the notion of quasi-needs, social-needs and deficit motivations. These needs are acquired motives that give rise to negative affect when unsatisfied (e.g., need for power, social status, achievement). Negative affect is reduced and goes toward a neutral state once such a need is satisfied, terminating the goal. This reduction in negative affect is itself reinforcing, and demonstrates the importance of implementing the capacity for relativistic subjective states. Stance is determined by whether pleasure or relief of pain is the guiding motivation. While more complex, this conceptualization permits acquisition of an endless array of motives and goals without the need to hardwire them as drives. It also more closely resembles human functioning.

Our early experience with this model suggests that establishing exact homeostatic set points and bounds is not critical to system functioning. Attaching negative affect or pain to homeostatic imbalances creates temporary drive states that motivate regulatory behavior. We have observed that the same behavior results regardless of the values established. Instead, the temporal cycles for satisfying drives vary with differences in the strength of affect arising from imbalances, resulting in behavioral differences comparable to temperament observed in humans. Unless the system is placed in an environment where satisfaction of imbalances is impossible, extremes are never reached, obviating the need for boundaries. Our system can function without predetermining "correct" set points or boundaries because the system's emotional behavior is not defined on the basis of its distinct homeostatic drive states, as it is in Breazeal's (2001) model.

When emotion arises as a fixed consequence of cognition or of some appraised environmental event, affect does not guide or influence cognition but is determined by it. We believe flexible thought can be achieved by linking affect to semantic meaning and using affect as a weighting mechanism, a significance indicator, a tuning mechanism for attention and memory, a choice mechanism, and a motivator of situation-appropriate behavior linked to accomplishing desired goals. This potential for flexibility is diminished when emotionality is specified to a system, not emergent from it.

References

Breazeal, C. 2001. *Designing Sociable Machines*. The MIT Press. Forthcoming.

Brooks, R. A. 1986. Achieving Artificial Intelligence through Building Robots. MIT AI Lab Memo 899.

Collins, A. & Loftus, E. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407-428.

Panksepp, J. 1998. *Affective Neuroscience*. Oxford Press.

Quillian, M. 1966. *Semantic Memory*. Cambridge, MA: Bolt, Beranak and Newman.

Reeve, J. 1997. *Understanding Motivation and Emotion*, Second Edition. Harcourt Brace College Publishers.

Shibata, T., T. Tashima, & K. Tanie 1999. Emergence of Emotional Behavior through Physical Interaction between Human and Robot, *Proceedings of the 1999 IEEE*

International Conference on Robotics and Automation (ICRA'99), Vol. 4, Page(s): 2868 –2873.