Cognitive Robotics
AAAI Technical Report WS-12-06

An Approach to Building Emotional Intelligence in Artifacts

Alexei V. Samsonovich
Krasnow Institute for Advanced Study, George Mason University
4400 University Drive MS 2A1, Fairfax, VA 22030-4444, USA
asamsono@gmu.edu

Abstract

Artificial emotional intelligence is vital for integration of future robots into the human society. This work introduces one possible approach to representation and processing of emotional mental states and attitudes (appraisals) in a cognitive architecture. The developed framework will allow for implementation of emotional intelligence in artifacts, including emotionally informed behavior and self-regulation, recognition of emotional motives in actions of other agents, and natural emergence of social emotional attitudes, interpreted here as higher-order appraisals. The proposed validation is based on a simulation of emergence of emotional relationships in a small group of agents.

Keywords: emotional intelligence; biologically inspired cognitive architectures; social emotions; human-level AI

Introduction

Recent years presented us with an explosion of research on computational models of emotions and affective computing (Picard, 1997; Hudlicka, 2011). Yet, there is no generally accepted complete, unifying theory of emotions that could take form of a computational model to be used in human-level emotional artificial intelligence. Even the very word "emotion" is not generally accepted as a well-understood scientific term. Different schools sometimes speak different languages and maintain very different views. The consensus is limited to the level of a basic model (Osgood et al. 1957; Russell 1980) known by different names overviewed below. At the same time, there is no general consensus on the computational understanding of even basic emotions like joy or fear, and in this situation, higher human emotions (usually referred to as "complex" or "social" emotions in the literature) inevitably escape attention in the fields of artificial intelligence and cognitive modeling. Nevertheless, their computational understanding is vital for the development of virtual agents and robots capable of adapting to the real world and working together with humans as partners (Buchsbaum et al. 2004; Parisi and Petrosino 2010).

Artificial emotional intelligence could be the key to a human-level artificial intelligence in general. From this perspective, adding advanced emotional capabilities to **cognitive architectures** (Gray 2007) is a critical milestone on the roadmap to human-level artificial intelligence. At the same time, the necessary for this step underlying theoretical and computational framework is missing.

In general, the notion of subjective emotional feeling is problematic in modern science: problems trace back to the general problems associated with the notion of consciousness (Chalmers 1996). At the same time, there is no need to solve the "hard problem" before describing psychological and functionalist aspects of emotions mathematically and using this description as a basis to replicate the same features in artifacts.

The purpose of the present study is to develop a simple computational framework capturing the functionality of emotions in general, including "complex" and "social" emotions. This framework is expected to result in development of new capabilities in artifacts, including automated recognition of higher-level emotional motivations in agent behavior, emotionally informed metacognition, and more.

Theoretical Background

Overview of the State of the Art

Research literature on emotions is immense, while there is no generally accepted unifying theory, and the prospect for developing it in the near future is bleak. While most modern studies of emotions do not go beyond phenomenology, a scientific theory must introduce general principles or mechanisms explaining the nature of the phenomenon that could be tested experimentally. Currently, a number of views and theories exist that attempt to relate emotions to first principles and/or to experimental data in neurophysiology, psychology, psychiatry, sociology, theory of evolution, theory of information, control theory, and beyond. A few of them are briefly outlined below.

Neurophysiologically, emotional reactions are associated with certain brain structures (such as nucleus accumbens, anterior cingulate and orbitofrontal cortices, the amygdala, hypothalamus, ventral tegmental area, etc.) and certain neurotransmitters. E.g., dopamine release in nucleus accumbens results in a feeling of pleasure and is responsible for the development of emotional memories, e.g., leading to drug addiction (Olds 1956; Wise 1978; Kringelbach 2009). Neurophysiological constraints like these cannot be ignored in construction of models of emotional intelligence, yet they alone are not sufficient.

Psychological and computational models of emotions attempt to reduce a large variety of affects, appraisals, emotions, feelings, moods, traits, attitudes, preferences, etc.[1] to a few universal constructs. Here, the main kinds of approaches are (Hudlicka 2011): (i) taxonomies, (ii) dimensional models, examples of which are the semantic differential model (Osgood et al. 1957) with its variations known by different names (EPA: evaluation, potency, arousal, PAD: pleasure, arousal, domainance, Circumplex: Russell 1980, etc.); and (iii) cognitive component models, or appraisal models, the most popular of which is currently OCC (Ortony, Clore and Collins 1988) because of its suitability for computer implementations (Steunebrink, Dastani, and Meyer 2007). The idea of OCC is to define all special circumstances under which emotions emerge, and also how emotion intensities change over time in those cases. However, this model is merely phenomenological, with the source of information being human intuition. Many alternative computational models of emotions are conceptually similar (e.g., Castelfranchi and Miceli 2009). Applications include modeling of relations development in social networks (e.g., trust relations: Sabater et al. 2006). A review of models of emotions cannot be provided here.

The simple bottom line is that, fortunately, most of the basic emotions, affects, feelings, etc.[1] can be efficiently characterized by a small number of very general semantic characteristics: attributes of emotions, that may include valence, dominance, arousal, surprise, etc. Together they form an attribute space, or "semantic map" of emotions that is known in many variations under different names (e.g., EPA, PAD), while in fact most of these models map onto each other. Adding more characteristics allows one to build more accurate representations (sliding to the level of OCC-like models), while selecting fewer characteristics allows one to build simpler, more parsimonious models.

[1] It is impossible here to discuss various interpretations of these terms and differences among them; therefore, "emotion" is used as a generic word.

Preliminary Observations

Yet, something is missing in this picture. E.g., at the level of EPA, it may be hard to define a boundary between basic emotions like fear or anger and complex emotions like jealousy or shame. It seems that complex, or social emotions must be in principle distinct from basic emotions not in their EPA values, but because they involve an element that is not present in basic emotions, which makes them "complex" or "social": for instance, this could be an element of metacognition.

A related observation: typically in affective modeling, emotional and cognitive components are added to a cognitive architecture separately. Frequently, their integration appears to be a challenge. By contrast, in humans, cognition and emotions are intimately mixed from birth and develop inseparably (Phelps 2006).

In general, despite the large amount of research literature on emotion modeling, the science of emotions appears to be at its infancy: it is limited to phenomenology, and even at this level there is no general consensus. In development of a theory based on phenomenology, the first step should consist in analysis of fundamental logical possibilities. The next step would be to choose parsimonious building blocks of the theory: in our case, abstract emotional constructs that can be implemented and studied computationally. Finally, the task would be to relate those primary as well as emergent in simulations constructs to known examples of emotions in real life. An attempt to make a move in this direction is undertaken in the present study. In order to proceed with it, two concepts need to be described first.

Mental State Formalism

The mental state formalism used here (Samsonovich, De Jong and Kitsantas, 2009) was developed in connection with the GMU BICA (Samsonovich and De Jong 2005) and Constructor (Samsonovich 2009) cognitive architectures based on previous ideas (Samsonovich and Nadel 2005). Related yet different frameworks were independently described in the literature (e.g., McCarthy, 1993; Nichols and Stich, 2003; Scally, Cassimatis and Uchida, 2012).

The essence of the mental state formalism is in the observation that cognitive representations in human working memory are usually attributed to some instance of a Self of some agent. This attribution together with its governing laws appears to be the only mechanism that instantiates the Self of the subject in the brain (with other selves possibly also represented at the same time in the same working memory). According to this attribution, all higher-level symbolic representations in working memory of a cognitive architecture can be partitioned into "boxes" labeled in an intuitive self-explanatory manner: "I-Now", "I-Previous", "He-Now", and so on, accordingly to the represented mental perspectives (Figure 1).

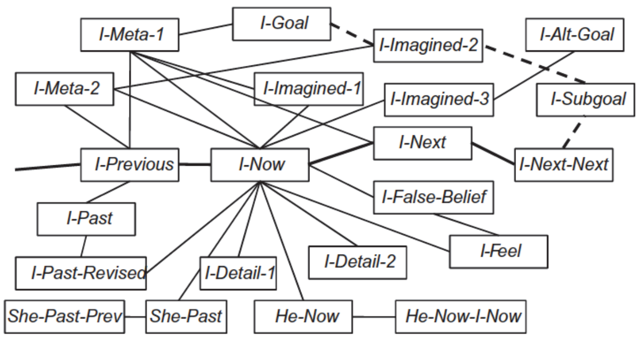

Figure 1. *Examples of possible mental state labels (boxes) and relations (lines) in working memory. From Samsonovich et al. (2009). Each box contains elements of awareness (not shown).*

These boxes represent mental states. Each "box" is filled with elements of awareness (not shown in Figure 1). In GMU BICA and Constructor these elements are instantiated using schemas[2] that may represent objects, relations, features, events, etc. The box label characterizes the mental perspective of the subject who is aware of them.

A **mental state** can be defined as the content of awareness of a subject that is associated with one mental perspective (Samsonovich et al., 2009). This mental perspective is efficiently captured by the self-explanatory mental state label (e.g., "I-Now"). The entire diagram (Figure 1) represents a state of working memory of one agent. Episodic memory (not shown) consists of similar structures that are "frozen" in long-term memory.

For the present purposes we shall stay at a high level of abstraction, ignoring the contents of mental states as much as possible, and will focus on relations of mental states to each other. Contents of interest here are emotional states and attitudes that represent subject's feelings about self and others, as well as about objects, relations, and about other feelings. The term "attitude" here stands for an attribute of a schema (Samsonovich et al. 2006). Example: an apple (schema) may have an attitude "imagined, desired".

Weak Semantic Cognitive Mapping

The term "weak semantic cognitive mapping" was described by Samsonovich, Goldin and Ascoli (2010). In general, the idea of semantic cognitive mapping is to allocate representations (e.g., words) in an abstract space based on their semantics. This paradigm is common for a large number of techniques, from the latent semantic analysis (LSA) to circumplex models. Traditionally, the metrics that determines allocation of symbols in space is a function of their semantic dissimilarity. In contrast, the idea of **weak** semantic cognitive mapping is not to separate all different meanings from each other, but to arrange them based on the very few principal semantic dimensions. These dimensions emerge automatically, if the strategy is to pull synonyms together and antonyms apart (Samsonovich and Ascoli 2007, 2010).

The beauty of this approach is that nobody needs to define semantic features of the space *a priori*: they are created by energy minimization, and their semantics are defined by the entire distribution of representations on the map. Most experts who first see weak semantic cognitive maps cannot understand the difference between this approach and LSA. However, these are two completely different approaches. In traditional techniques based on LSA (Landauer et al. 2007), the starting point is a feature space, where dimensions have definite semantics a priori. In weak semantic mapping, the space coordinates have no semantics associated with them a priori: instead, words are allocated **randomly** in a multidimensional space. Then an energy minimization process starts that pulls synonym vectors together and antonym vectors apart. The optimized distribution of words has emergent semantics of its dimensions. Only then the principal component analysis is used to reveal the main emergent semantic dimensions of the map (Samsonovich and Ascoli 2007, 2010).

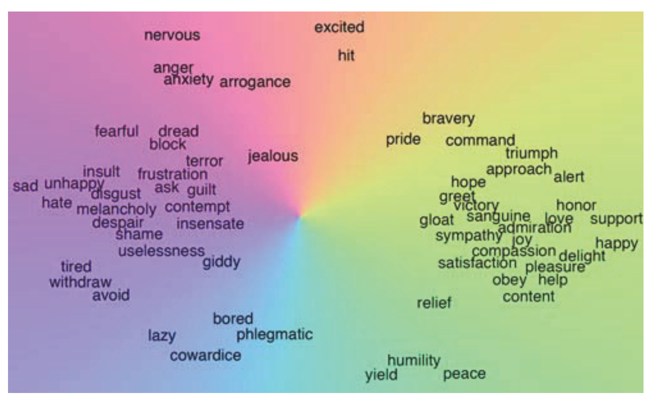

Figure 2. *A sample from the weak semantic cognitive map described by Samsonovich and Ascoli (2010). The "synesthetic" color enhancement follows the ideas of Michael Sellers. Colors represent, green: PC1 (pleasure), red: PC2 (dominance/arousal). The sum of RGB values is fixed.*

The map a part of which is shown in Figure 2 includes 15,783 words and was constructed based on the dictionary of English synonyms and antonyms available as a part of Microsoft Word (Samsonovich and Ascoli, 2010; a similar

[2] The term "schema" here has a specific meaning, different from its meanings in psychology or computer science (Samsonovich et al. 2006).

map was also constructed using WordNet in the same work). This map does not separate well different meanings from each other: e.g., basic and complex feelings. However, it classifies meanings consistently with their semantics. Figure 2 represents the first two principal components (PC) of the distribution. The axes of the map are defined by the PCs. In a very approximate sense, PC1 (the horizontal dimension in Figure 2, coded by the green-magenta gradient) can be associated with valence, positivity, pleasure, attractiveness, while PC2 combines the notions of dominance, arousal, potency, strength (the vertical dimension in Figure 2, coded by red-cyan).

In the following sections we will need the actual map coordinates for some of the word examples. They are given in Table 1. Normalized values of PC1 and PC2 are taken from the same map as shown in Figure 2, available as part of the materials of Samsonovich and Ascoli (2010).

Table 1. *Weak semantic cognitive map coordinates for actions (values taken from materials of Samsonovich and Ascoli 2010).*

	PC1	PC2
hit	0.26	1.07
yield	0.43	-1.03
greet	0.93	0.15
withdraw	-1.66	-0.42

Introducing Building Blocks

Basic Ideas

As always in science, it makes sense to start with the most basic questions and simplest possible answers. The first question is, how to choose the simplest possible general elements representing emotions that can be added locally to any system of mental states, without destroying the formalism? There are a few possibilities here. E.g., one can add an emotional characteristic to a mental state, or to an element of a mental state – an instance of a schema in it. This gives us two new elements that extend the formalism. In the first case, the new global characteristic assigned to a mental state will be called an **emotional state** (Figure 3 A). In the second case, the characteristic assigned to an instance of a schema will be called an emotional attitude, or **appraisal**[3] (Figure 3 B, C). This schema may represent another agent: then it will link two mental states.

[3] The term "attitude" in the mental state formalism has a more general and possibly counterintuitive meaning (Samsonovich et al., 2009). The more pertinent to emotions term "appraisal" will be used here in order to refer to emotional attitudes.

Consider an example (Figure 3 C): I can be aware of another agent A represented by a schema of A in my I-Now. This schema also serves as a link to A-Now: my mental simulation of the mind of A. An instance of my appraisal of A is attached as an attribute to the schema in I-Now. It tells me (the subject) how I emotionally evaluate that agent. For example, when I honor A, or when I submit to A, my appraisal of A has positive dominance (Figure 3 C). At the same time, if I take an action to yield to A, then my appraisal of the yield action in I-Now has a negative dominance (Table 1), while my appraisal of A increases its dominance due to the performed action.

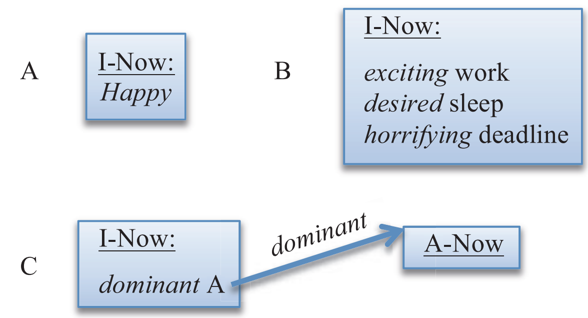

Figure 3. *A-C: Examples of simplest emotional elements in the mental state formalism. Boxes represent mental states. Mental state labels are underlined, emotional characteristics are italicized. A: an example of an emotional state; B: appraisals of processes and events; C: appraisal of an agent. The arrow goes from the subject who is apprising to the appraised target.*

As to the nature of values that the new characteristics can take, the parsimonious choice made here is to use vectors from the weak semantic cognitive map as values for all emotional characteristics, including emotional states and appraisals of animate and inanimate elements.

Understanding "Complex", or "Social" Emotions

The next level of complexity of building blocks in this framework must be represented by feelings about feelings: this is the remaining possibility. These elements are by definition higher-order feelings, and therefore they will be represented as appraisals of appraisals, or **higher-order appraisals** (HOA). The general building block that is used to create them is a schema called here a **HOA schema**.

It appears that in many cases higher-order feelings map nicely on what is known in the literature as complex, or social emotions: e.g., envy, jealousy (Parrott 2001). It is tempting to represent them by schemas. Related ideas were discussed recently in an attempt to explain the unity of social emotions (Castelfranchi and Miceli 2009). Figure 4 illustrates this possibility using one and the same abstract HOA schema in four contexts with different outcomes.

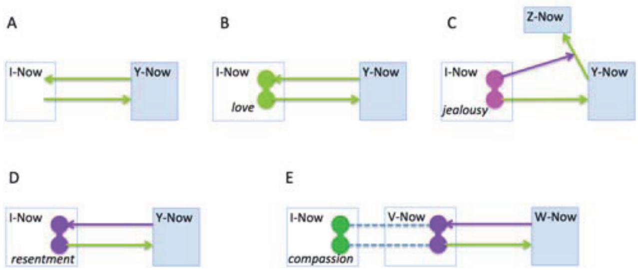

Figure 4. *Examples of abstract higher-order appraisals (HOA) that can emerge in working memory as different states of one and the same HOA schema (the filled "figure 8"). An interpretation relates them to familiar notions like jealousy (Parrott 2001) that develops in the sequence A, B, C. Colors correspond to Figure 2.*

Examples of Figure 4 illustrate a possibility that one HOA schema may play different roles in different contexts, and therefore account for a family of related social emotions rather than just one of them. In this case, the framework could predict the emergence of new "abstract" feelings that can be then identified with familiar real-life stereotypes. This could lead to an understanding of the mechanisms of complex, or social emotions like shame, pride, resentment, jealousy, compassion, humor, etc. While the labeling in Figure 4 remains a speculation, the above consideration supports the view that "complex", or "social" emotions in principle can be understood based on HOA.

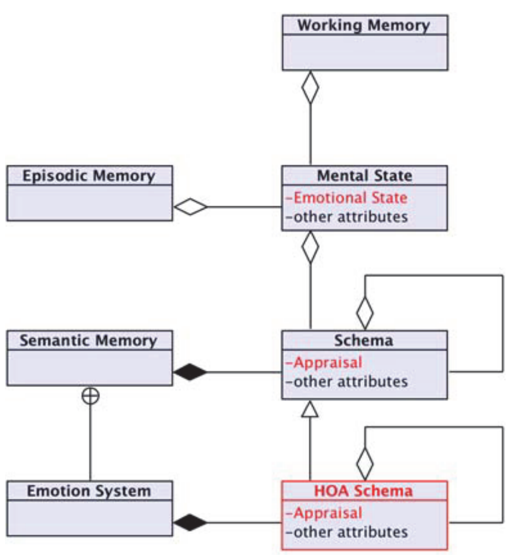

Figure 5. *Essential core UML class diagram of the Constructor cognitive architecture extended to enable emotional intelligence. Emotional elements (shown in red) include two attributes (Emotional State, Appraisal) and higher-order appraisal (HOA) schemas. Appraisal values are shown in Figure 2. For details of the architecture, see (Samsonovich et al. 2006, Samsonovich 2009).*

Putting the Framework Together

In summary, the proposed emotional extension of the formalism of schemas and mental states that underlies aforementioned cognitive architectures is based on three new elements:

- emotional states;
- first-order appraisals: appraisals of objects, facts, events, actions, relations, etc.; as well as appraisals of agent minds and personalities;
- appraisals of appraisals, or HOA.

All these characteristics take values on the weak semantic cognitive map (or in the EPA space). As to their specific implementation, the first two can be added to the lists of standard attributes of mental states and schemas, whereas the third element can be added to the framework as a schema on its own (Figure 5). For details of GMU BICA and Constructor architectures, the Reader is referred to previous publications (Samsonovich and De Jong 2005, Samsonovich et al. 2006, 2009; Samsonovich 2009).

In addition to these elements, in order to complete the framework, laws of interactions between emotional and other elements need to be specified, including: (i) effects of events on emotional states and appraisals; (ii) effects of emotional states and appraisals on cognitive processes, e.g., selection of actions. This is done in a simple illustrative example in the following section.

A Simple Computational Illustration

The paradigm includes a group of N agents embedded in some virtual environment (Figure 6 A). For the purpose of a simplest example we shall ignore any environmental (including spatial) aspects of the problem and assume that agents interact continuously regardless of where they are. They interact by spontaneously performing actions from a pre-defined repertoire. The repertoire includes four possible actions: hit, yield, greet, or withdraw. All of these actions are appraised by the agents (the action appraisal values are fixed and given in Table 1). The goal in this exercise is to see what stable patterns of emotional relationships among agents will emerge, and how the outcome may depend on model parameters. The meta-goal, however, is to better articulate the model and show it at work, albeit in a trivial example. There is only one working memory in this extremely simplified multi-agent cognitive architecture. This generic working memory is presumed to correctly describe working memory of any of the participating agents: there are no false beliefs or subjective biases. This also assumes that all agents receive the same information about the world and always "correctly interpret" each other's motivations. The obvious indexical differences between the N agent perspectives are easily taken into account in derivation of the dynamic equations.

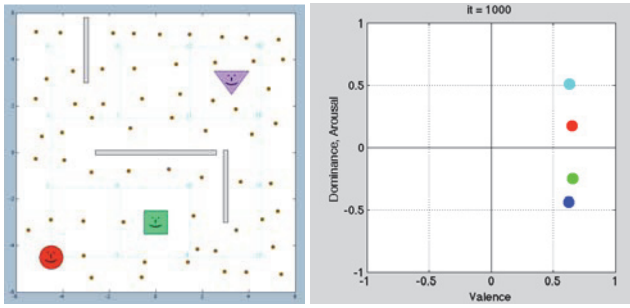

Figure 6. *Left: The simulation paradigm. Agents are embedded in some virtual environment. Right: A typical configuration with N = 4 agents on the weak semantic map. Colors are arbitrary.*

There are N mental states in working memory, each corresponding to the state of awareness of one agent taken at the present moment. The appraisal of a given mental state is the same for all appraisers in this model: this justifies using one working memory. Emotional states and higher-order appraisals are not simulated. Therefore, there are only N dynamic emotional characteristics in this model: appraisals of the N agents. They are initiated to very small random values at the beginning of the simulation epoch, in order to break the symmetry. Each of the four possible actions has a fixed appraisal given in Table 1. Appraisal values A are 2-D vectors that are treated here for convenience of implementation as complex numbers:

$A = (valence, dominance)$. In this case,
valence = Re(A), and *dominance* = Im(A).

The simulation epoch consists of a sequence of iterations performed at each moment of discrete time t. One iteration includes the following essential steps: (i) compute action probabilities, (ii) select and perform action, (iii) update appraisals of the actor and the target of action. Dynamical equations used to update the appraisals are:

$$A^{t+1}_{target} = (1-r)A^{t}_{target} + rA_{action} \qquad (1)$$
$$A^{t+1}_{actor} = (1-r)A^{t}_{actor} + rA^{*}_{action}$$

Here t is the moment of discrete time, r is a small positive number (a model parameter that was set to 0.01). The likelihood L of a specific action is proportional to

$$L_{action} \sim \left[\mathrm{Re}\left(A_{action}\left(A^{*}_{actor} + A_{target}\right)\right)\right]_{+}.$$

Here $[x]_{+}$ is equal to the positive values of x and is zero otherwise, A^{*} is the complex conjugate of A. Intuitively, this formula means that the action is more likely to be selected, when its appraisal matches the appraisal of the actor and also matches the appraisal of the target, in which the dominance component is inverted.

This model is easy to solve approximately analytically: first, in a reasonably good approximation, the two dimensions become independent; then, an observation is made that dynamics are close to iterative multiplication of a matrix, which results in singling out its main eigenvector (details will not be given here). This analytical prediction is consistent with the simulation results.

Results

It is found that stable patterns of mutual appraisals (that correspond to configurations of emotional relationships among agents) develop in this model in ~100 iterations. E.g., with the choice of parameters specified above, a pattern always develops in which all N appraisals have positive valence. At a small N, each actor reaches its stationary position in ~100 iterations. At large N, the final configuration remains stationary only macroscopically (as a "cloud"), while there is no microscopically stationary configuration. The qualitative outcome for N=2 is nearly obvious based on the above analysis. In a stationary configuration, the two agent vectors tend to be complex conjugates of each other with the positive real part. At N=3 the stable configuration remains qualitatively the same, while the third actor takes position in the middle between the two, at zero dominance (analytical and numerical result). At $N \geq 4$ the configuration is microscopically undetermined: actors do not have permanent stationary positions in the cloud. They tend to spread uniformly in a vertical line at a positive valence, and keep drifting up and down, switching positions with each other.

Now it is interesting to see what happens when HOA are added to dynamics of the system. In the language of this simple model, adding HOA schemas that have effect on dynamics means adding nonlinear in A terms to (1). This can stabilize microscopic configurations above N=3: in other words, permanent relations in a large population of agents may be expected to emerge in the system. The study will be continued and presented elsewhere.

Discussion

Traditionally, emotional information processing in cognitive modeling is contrasted with rational cognition, and it is considered a challenge to put the two together in one cognitive architecture. By contrast, one basic idea underlying this study is that cognitive architectures should be designed in such manner that all information processing in them could be regarded as "emotional". In particular, this means that (i) goals should originate from intrinsic emotions rather than from externally given instructions; (ii) emotional components should be essential to any part of the cognitive process in the architecture; and (iii) the outcome of each cognitive process should be captured by a certain emotional state. From this point of view, it would be misleading to think that emotions, moods and feeling may only subserve impulsive responses or biases, whereas rational planning and decision making is emotion-independent. Instead, emotional elements should find their proper place in all basic mechanisms of cognition in future cognitive architectures. But this can only happen if the right approach can be found to implement and use them.

Do Artifacts Need Emotional Intelligence?

The simulated example paradigm was inspired by the paradigm used by Heider and Simmel (Figure 7) in their behavioral study of human subjects (Heider and Simmel, 1944). Besides its main purpose to serve as a test for humans, their example clearly demonstrates that in order to be emotionally understood and accepted by humans on an equal footing, artifacts do not need to be human-level realistic in their appearance, or in their ability to control voice and motion. Regardless of physical abilities, they need to demonstrate a human-level emotional competency, and therefore they need to be emotionally intelligent at a human level. The same conclusion follows from research on the sense of presence in virtual environments (Herrera et al., 2006). On the other hand, research on human learning tells us that emotions play a vital role in it (Zimmerman 2008), and therefore it seems that it is necessary for artifacts to have human-level emotional intelligence to be able to learn like humans.

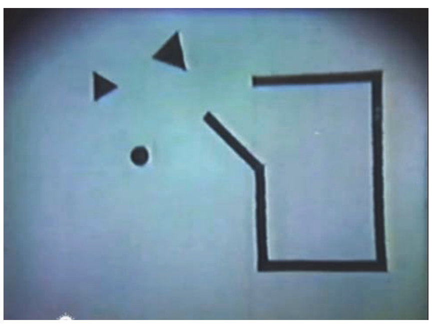

Figure 7. *A frame from the animation by Heider and Simmel (available at http://www.youtube.com/watch?v=76p64j3H1Ng) used in their study (Heider and Simmel, 1944). The three small shapes are the actors that create an impression of emotional relationships among them by simply moving around the square.*

Emotions in Mainstream Cognitive Architectures

Soar (Laird et al., 1987) and ACT-R (Anderson and Lebiere 1998) are the two most widely known and used cognitive architectures. There were recently numerous works on implementation and study of emotions in Soar and ACT-R, along with other nowadays popular features like episodic memory, which is not discussed here. It is relatively easy to implement some aspect of emotions in a cognitive architecture. But this does not necessarily solve the problem. For example, the recently extended version of Soar (Laird 2008) implements an appraisal theory of Scherer (that belongs to the same family as OCC) in its Appraisal Detector, which is used for reinforcement learning. A limitation here is that appraisal is evaluated as a global characteristic of the situation of the agent.

Potential Applications

Today's large volumes of surveillance data pose a challenge of extracting vital information from the data automatically. The task may remain unsolved even after relevant objects, features and apparent relations have been identified and represented in a symbolic format. In many important cases, the remaining task involves higher cognitive analysis of human mental states (elements in human Theory of Mind), which today requires a human analyst.

Biologically-inspired affective cognitive architectures increasingly attract attention in this context as an efficient, robust and flexible solution to the challenge of real-world sensory data analysis. At the same time, modeling of higher cognition in cognitive architectures is often limited to simplistic algorithms and separated from biologically inspired information processing.

Examples of potential applications of future intelligent agents that will be based on this approach include: (i) detection, anticipation and control of ad hoc social interactions related to spontaneous violence, accidents or a natural disaster, (ii) detection of individual and group activities that are likely to be related to terrorism, (iii) prediction of a potentially harmful individual condition: e.g., psychological breakdown of a human operator, driver, or air traffic controller.

Metacognitive Theory-of-Mind skills that a human analyst must have in order to be efficient in the above tasks include the abilities to infer, attribute, simulate and predict mental states of actors involved in the ongoing action. The approach developed in this work supports the view that an artifact can and should have similar metacognitive abilities in order to be human-level intelligent and successful in solving tasks of the mental state analysis.

Acknowledgments

I am grateful to Dr. Michael Sellers and many others for discussions of the status of the field of emotional artificial intelligence (some of these discussions are available online at http://bicasociety.org/videos/vp.html). I am very grateful to Dr. Kenneth De Jong for early discussions of the ideas underlying this study. I am mostly grateful to Dr. Claudio Cioffi-Revilla for proofreading and very useful comments on an early version of the manuscript. Finally, I would like to thank the three anonymous reviewers for their constructive criticism and suggestions, and the Workshop Chairs for giving me the opportunity to present this work.

References

Anderson, J. R. and Lebiere, C. 1998. *The Atomic Components of Thought.* Mahwah: Lawrence Erlbaum Associates.

Buchsbaum, D., Blumberg, B., and Breazeal, C. 2004. Social learning in humans, animals and agents. *AAAI Fall Symposium Technical Report* FS-04-05, pp. 9-16. Menlo Park, CA: The AAAI Press.

Castelfranchi, C., and Miceli, M. 2009. The cognitive-motivational compound of emotional experience. *Emotion Review* 1 (3): 223-231.

Chalmers, D.J. 1996. *Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford University Press.

Gray, W. D. (Ed.) 2007. *Integrated Models of Cognitive Systems. Series on Cognitive Models and Architectures.* Oxford, UK: Oxford University Press.

Heider, F, and Simmel, M. 1944. An experimental study of apparent behavior. *American Journal of Psychology* 57: 243–259.

Herrera, G., Jordan, R., & Vera, L. 2006. Agency and presence: A common dependence on subjectivity? *Presence: Teleoperators and Virtual Environments* 15 (5): 539-552.

Hudlicka, E. 2011. Guidelines for designing computational models of emotions. *International Journal of Synthetic Emotions*, 2 (1): 26-79.

Kringelbach, M.L. 2009. *The pleasure center: Trust Your Animal Instincts.* Oxford University Press.

Laird, J.E., Newell, A., and Rosenbloom, P.S., 1987. SOAR: An architecture for general intelligence. *Artificial Intelligence* 33: 1-64.

Laird, J.E. 2008. Extending the Soar cognitive architecture. In P. Wang, B. Goertzel and S. Franklin, eds. *Artificial General Intelligence 2008: Proceedings of the First AGI Conference,* pp. 224-235. Amsterdam, The Netherlands: IOS Press.

Landauer, T.K., McNamara, D.S., Dennis, S., and Kintsch, W., (eds.). 2007. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

McCarthy, J. 1993. Notes on formalizing context. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 555–562. Menlo Park, CA: AAAI Press.

Nichols, S. and Stich, S. 2003. *Mindreading: An Intergrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

Olds, J. 1956. Pleasure centers in the brain. *Scientific American*, 105-116.

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

Osgood, C.E.; Suci, G.; Tannenbaum, P. *The Measurement of Meaning.* University of Illinois Press: Urbana, IL, 1957.

Parisi, D. and Petrosino, G. 2010. Robots that have emotions. *Adaptive Behavior* 18 (6): 453-469.

Parrott, W. G. 2001. Experiences of envy and jealousy. In Parrott, W. G. (Ed.). *Emotions in Social Psychology,* pp. 306-320. Philadelphia, PA: Psychology Press.

Phelps, E.A. 2006. Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology* 57: 27-53.

Picard, R. 1997. *Affective Computing.* Cambridge, MA: The MIT Press.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6): 1161-1178.

Sabater, J., Paolucci, M., and Conte, R. 2006. Repage: Reputation and image among limited autonomous partners, *Journal of Artificial Societies and Social Simulation* 9 (2): 3.

Samsonovich, A. V. 2009. The Constructor metacognitive architecture. In Samsonovich, A. V. (Ed.). *Biologically Inspired Cognitive Architectures II: Papers from the AAAI Fall Symposium. AAAI Technical Report* FS-09-01, pp. 124-134. Menlo Park, CA: AAAI Press.

Samsonovich, A. V. and Ascoli, G. A. 2007. Cognitive map dimensions of the human value system extracted from natural language. In Goertzel, B. and Wang, P. (Eds.). *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. Proceedings of the AGI Workshop 2006. Frontiers in Artificial Intelligence and Applications,* vol. 157, pp. 111-124. IOS Press: Amsterdam, The Netherlands. ISBN 978-1-58603-758-1.

Samsonovich, A. V. and Ascoli, G. A. 2010. Principal Semantic Components of Language and the Measurement of Meaning. *PLoS ONE* 5 (6): e10921.1-e10921.17.

Samsonovich, A. V. and De Jong, K. A. 2005. Designing a self-aware neuromorphic hybrid. In K.R. Thorisson, H. Vilhjalmsson, and S. Marsela (Eds.). *AAAI-05 Workshop on Modular Construction of Human-Like Intelligence: AAAI Technical Report*, volume WS-05-08, pp. 71–78. Menlo Park, CA: AAAI Press.

Samsonovich, A. V. and Nadel, L. 2005. Fundamental principles and mechanisms of the conscious self. *Cortex* 41 (5): 669–689.

Samsonovich, A. V., and Jóhannsdóttir, K. R. (Eds.). 2011. *Biologically Inspired Cognitive Architectures 2011: Proceedings of the Second Annual Meeting of the BICA Society. Frontiers in Artificial Intelligence and Applications,* vol. 233. Amsterdam: IOS Press. ISSN 0922-6389.

Samsonovich, A. V., Ascoli, G. A., De Jong, K. A., and Coletti, M. A. 2006. Integrated hybrid cognitive architecture for a virtual roboscout. In M. Beetz, K. Rajan, M. Thielscher, and R.B. Rusu (Eds.). *Cognitive Robotics: Papers from the AAAI Workshop, AAAI Technical Reports* WS-06-03, pp. 129–134. Menlo Park, CA: AAAI Press.

Samsonovich, A. V., De Jong, K. A., and Kitsantas, A. 2009. The mental state formalism of GMU-BICA. *International Journal of Machine Consciousness* 1 (1): 111-130.

Samsonovich, A. V., Goldin, R. F., and Ascoli, G. A. 2010. Toward a semantic general theory of everything. *Complexity* 15 (4): 12-18.

Samsonovich, A. V., Jóhannsdóttir, K. R., Chella, A., and Goertzel, B. (Eds.). 2010. *Biologically Inspired Cognitive Architectures 2010: Proceedings of the First Annual Meeting of the BICA Society. Frontiers in Artificial Intelligence and Applications,* vol. 221. Amsterdam: IOS Press. ISSN 0922-6389.

Scally, J.R., Cassimatis, N.L., and Uchida, H. 2012. Worlds as a unifying element of knowledge representation. *Biologically Inspired Cognitive Architectures*, 1 (in press).

Steunebrink, B. R., Dastani, M., and Meyer, J. -. C. 2007. A logic of emotions for intelligent agents. *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007), pp. 142-147.* Menlo Park, CA: AAAI Press.

Wise, R.A. 1978. The role of reward pathways in the development of drug dependence. *Pharmacology and Therapeutics* 35(1-2):227-63.

Zimmerman, B.J. 2008, Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal* 45 (1): 166-183.