Human-like Emotional Responses in a Simplified Independent Core Observer Model System

David J. Kelley[1] and Mark R. Waser[1, 2]

[1] Artificial General Intelligence Inc., Kent, WA, USA
[2] Digital Wisdom Institute, Richmond, VA USA
david@artificialgeneralintelligence.com
mark.waser@wisdom.digital

Abstract. Most artificial general intelligence (AGI) system developers have been focused upon intelligence (the ability to achieve goals, perform tasks or solve problems) rather than motivation (*why* the system does what it does). As a result, most AGIs have an unhuman-like, and arguably dangerous, top-down hierarchical goal structure as the sole driver of their choices and actions. On the other hand, the independent core observer model (ICOM) was specifically designed to have a human-like "emotional" motivational system. We report here on the most recent versions of and experiments upon our latest ICOM-based systems. We have moved from a partial implementation of the abstruse and overly complex Wilcox model of emotions to a more complete implementation of the simpler Plutchik model. We have seen responses that, at first glance, were surprising and seemingly illogical – but which mirror human responses and which make total sense when considered more fully in the context of surviving in the real world. For example, in "isolation studies", we find that any input, even pain, is preferred over having no input at all. We believe that the fact that the system generates such unexpected but "humanlike" behavior to be a very good sign that we are successfully capturing the essence of the only known operational motivational system.

Keywords: emotion, motivational system, safe AI.

1 Introduction

With the notable exception of the developmental robotics, most artificial general intelligence (AGI) system development to date has been focused more upon the details of intelligence rather than the motivational aspects of the systems (i.e. *why* the system does what it does). As a result, AGI has come to be dominated by systems designed to solve a wide variety of problems and/or to perform a wide variety of tasks under a wide variety of circumstances in a wide variety of environments – but with no clue of what to do with those abilities. In contrast, the independent core observer model (ICOM) [1] is designed to "solve or create human-like cognition in a software system sufficiently able to self-motivate, take independent action on that motivation and to further modify actions based on self-modified needs and desires over time." As a result, while most AGIs have an untested, and arguably dangerous, top-down hierarchical goal structure as their sole motivational driver, ICOM was specifically designed to have a human-like "emotional" motivational system that follows the 5 S's (Simple, Safe, Stable, Self-correcting and Sympathetic to current human thinking, intuition and feelings) [2].

Looking at the example of human beings [3-6], it is apparent that our decisions are not always based upon logic and that our core motivations arise from our feelings, emotions and desires – frequently without our conscious/rational mind even being aware of that fact. Damasio [7-8] describes how feeling and emotion are necessary to creating self and consciousness and it is clear that damage reducing emotional capabilities severely impacts decision-making [9] as well as frequently leading to acquired sociopathy whether caused by injury [10] or age-related dementia [11]. Clearly, it would be more consistent with human intelligence if our machine intelligences were implemented in the relatively well-understood cognitive state space of an emotional self rather than an unexplored one like unemotional and selfless "rationality".

While some might scoff at machines feeling pain or emotions or being conscious, Minsky [12] was clear in his opinion that "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions." Other researchers have presented compelling cases [13-16] for the probability of sophisticated self-aware machines necessarily having such feelings or analogues exact enough that any differences are likely irrelevant. There is also increasing evidence that emotions are critical to implementing human-like morality [17] with disgust being particularly important [18].

2 Methods

ICOM is focused on how a mind says to itself, "I exist – and here is how I feel about that". In its current form, it is not focused on the nuances of decomposing a given set of sensory input but really on what happens to that input after it's evaluated or 'comprehended' and ready to decide how 'it' (being an ICOM implementation) feels about it. Its thesis statement is that:

> *Regardless of the standard cognitive architecture used to produce the 'understanding' of a thing in context, the ICOM architecture supports assigning value to that context in a computer system that is self-modifying based on those value based assessments…*

As previously described [19], ICOM is at a fundamental level driven by the idea that the system is assigning emotional values to 'context' as it is perceived by the system to determine its own feelings. The ICOM core has both a primary/current/conscious and a secondary/subconscious emotional state -- each represented by a series of floating point values in the lab implementations. Both sets of states along with a needs hierarchy [20-21] are part of the core calculations for the core to process a single context tree.

Not wanting to reinvent the wheel, we have limited ourselves to existing emotional models. While the OCC model [22] has seemingly established itself as the standard model for machine emotion synthesis, it has the demonstrated [23] shortcoming of requiring intelligence before emotion becomes possible. Since the Willcox "Feelings Wheel" [24] seemed the most sophisticated and 'logical' emotion-first model, we started with that. Unfortunately, its 72 categories ultimately proved to be over-complex and descriptive rather than generative.

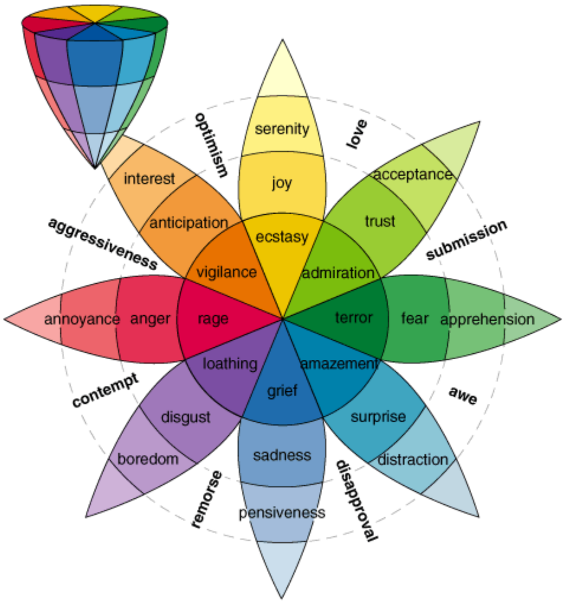

Fig. 1. The Plutchik model

The Plutchik model [25-27] starts with eight 'biologically primitive' emotions evolved in order to increase fitness and has been hailed [28] as "one of the most influential classification approaches for general emotional responses. Emotional Cognitive Theory [29] combines Plutchik's model with Carl Jung's Theory of Psychological Types and the Meyers-Briggs Personality Types.

3 Calculation

The default Core Context is the key elements pre-defined in the system when it starts for the first time. These are 'concept's that are understood by default and have predefined emotional context trees associated with them. They are used to associate emotional context to elements of context as they are passed into the core.

While all of these are hard coded into the research system at the start, they are only really defined in terms of other context being associated with them and in terms of emotional context associated with each element which is true of all elements of the system. Further these emotional structures or matrixes that can change and evolve over time as other context is associated with them. Some examples of these variables and their default values are:

- Action – The need to associate a predisposition for action as the system evolves.
- Input – A key context flag distinguishing internal imaginations vs external input.
- Pattern – A recognition of a pattern built-in to help guide context (based upon humans' inherent nature to see patterns in things).
- Paradox – A condition where 2 values that should be the same are not or that contradict each other.

Note that, while we might use these 'names' to make this item easily recognizable to human programmers, the actual internal meaning is only implied and enforced by the relationship of elements to other emotional values and each other and the emotional matrix used to apply those emotional relationships (i.e. we recognize that Harnad's grounding problem is very relevant).

The context emotional states and the states of the system are treated as 'sets' with matrix rules being applied at each cycle to a quickly-changing 'conscious' and a slower-moving 'subconscious' that more strongly tends towards default emotions. The interplay between them is the very heart of the system that creates the emotional subjective experience of the system.

```
$\forall\{\mathrm{E1, E3, \ldots, E72}\} \in Concious\ , E1 = Emotion1, E2 = \ Emotion2, \ldots, E72 = Emotions72$ ;
$\forall\{\mathrm{AE1, E3, \ldots, E72}\} \in Subconcious\ , E1 = Emotion1, E2 = \ Emotion2, \ldots, E72 = Emotions72$ ;

$\forall\ NewContext = f(\sum Inputs)\ or\ f(MemoryStack_n)$ ,
$\forall\ NewContext = fNeeds(NewContext)$ ,

    $\forall\{f\} \in ConciousRules\ \wedge \forall\{\mathrm{E1, E3, \ldots, E72}\} \in Concious\ , A = f(A \in Concious\ , \{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ ), B = f(B \in Concious\ , \{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ ), \ldots, D = f(D \in Concious\ , \{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ )$ ;
    $\forall\{f\} \in SubconciousRules\ \wedge \forall\{\mathrm{E1, E3, \ldots, E72}\} \in Subconcious\ , A = f(A \in Subconcious, \{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ ), B = f(B \in Subconcious\ , \{\mathrm{A, B, C, D}\} \in NewContext\ ), \ldots, D = f(D \in Subconcious\ , \{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ )$ ;
    $\forall\{f\} \in SubconciousRules\ \wedge \forall\{\mathrm{E1, E3, \ldots, E72}\} \in Concious\ , A = f(A \in Subconcious, \{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ ), B = f(B \in Subconcious\ , \{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ ), \ldots, D = f(D \in Subconcious\ , \{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ )$ ;
    $\forall\{f\} \in NewContextRules\ \wedge \forall\{\mathrm{E1, E3, \ldots, E72}\} \in NewContext\ , A = f(A \in NewContext\ , \{\mathrm{E1, E3, \ldots, E72}\} \in Concious\ ), B = f(B \in NewContext\ , \{\mathrm{E1, E3, \ldots, E72}\} \in Concious\ ), \ldots, D = f(D \in NewContext\ , \{\mathrm{E1, E3, \ldots, E72}\} \in Concious\ )$ ;
$\forall\ Action = fObserver(NewContext)$ ;

    $\forall\{N\} \in MemoryStack_n \quad = f\ (\mathrm{NewContext}, MemoryStack)$;
```

Fig. 2. Core Logic Notation/Pseudocode

New context associated with the object map or context tree of the current thought is executed against every single cycle regardless of whether its origin is external input or internal thoughts. Essentially the rules are then applied as to the relationships between those various elements which is after the needs and other adjustments to where it then falls into this final block which really is where the determination is made and it is in these rules applied here that we see the matrix of the system affecting the results of the isolation study.

4 Results

While investigating how the system behaved under a wide variety of circumstances, we encountered a series of cases whose results were initially very disturbing when testing what happened when we stopped all input (while ICOM continued to process how it felt) and then, finally, restarted the input. Imagine our surprise and initial dismay when the system, upon being presented only with pain and other negative stimulus upon the restarting of input, and actually "enjoyed" it. Of course, we should have expected this result. Further examination showed that the initial "conscious" reaction of ICOM was to "get upset" and to "desire" the input to stop – but that the "subconscious" level, the system "enjoyed" the input and that this eventually affected the "conscious" perception. This makes perfect sense because it is not that ICOM really "liked" the "pain" so much as it was that even "pain" is better than isolation – much like human children will prefer and even provoke negative reactions to avoid being ignored.

To be precise ICOM represents its internal emotional state by a set of ‘Plutchik’ models. Each model is a set of 8 floating point values that represent degrees of emotion. The system experiences ‘thoughts’ in the form of differential effects caused by matrix computations represented in the above functions from figure 2. That is to say if the system feels 5.3413523462455 of emotion E1 and the new context has the same E1 value of 5.3415453456544 then it is the difference in these that causes effect in the model.

These operations apply interests, needs and other underlying emotional factors to the context tree’s that are emotional structures of Plutchik models associated with the relational data structures that are effectively underlying thoughts that were surfaces to the core where the differences emotional factors affect the systems core and subconscious emotional states through those functions. Whereas this structure is designed to cause this emotional matrix structure to set its self-up in memory (while it is an misnomer it is easier to think of it as order ‘emerging’ out of the operations of the system as the system underlying awareness only exists in those emotional differentials that it experiences as it processed emotional context trees where as in this next diagram we see only a few of the values processed.

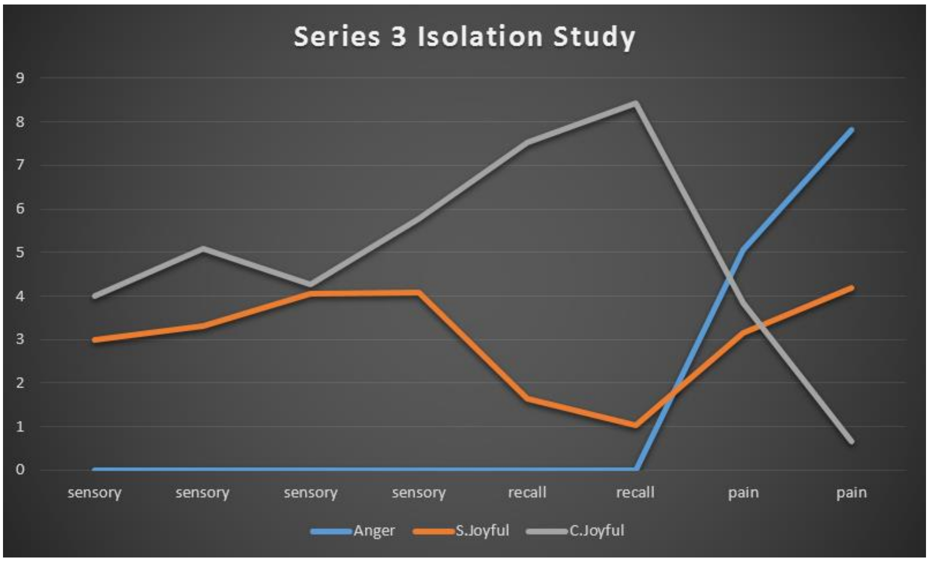

Fig. 3. Series 3 Isolation Study (x = input type w/time; y = intensity of emotion)

In the series 3 isolation study where the experiment was to place the system with random input and then isolate it from input and then provide pain we got results. While repeated with over 3+ million cycles the experiment consistently should similar results where the system became distraught in isolation once plain stimulus was applied through the test harness the system ‘conscious’ happy-ness dropped and anger sky rocketed what originally, we though anomalous was that the subconscious joy values rose dramatically.

5 Discussion

It’s always great when experiments produce unexpected emergent results that should have been anticipated because they are exhibited in the original system your model is based upon. We believe that the fact that the system spontaneously generates such unexpected but “humanlike” behavior to be a very good sign that we are successfully capturing the essence of the only known operational motivational system with a humanlike emotional “self”.

This kind of system with a subjective internal emotional landscape lays the ground work for some day creating systems that are sapient and sentience. It is in these systems with internal emotional landscape that experience subjective emotions from their standpoint we hope to create systems that can be taught ethical models that can our partners in the universe.

References

1. Kelley, D.: Self-Motivating Computation System Cognitive Architecture, (2016)
2. Waser, M.: Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence. In: AAAI Tech Report FS-08-04: Biologically Inspired Cognitive Architectures. http://www.aaai.org/Papers/Symposia/Fall/2008/FS-08-04/FS08-04-049.pdf
3. Haidt, J.: The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. Psychological Review 108, 814-823 (2001).
4. Minsky, M. L.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon & Schuster, New York (2006).
5. Hauser, M. et al: A Dissociation Between Moral Judgments and Justifications. Mind & Language 22(1), 1-27 (2007).
6. Camp, J.: Decisions Are Emotional, Not Logical: The Neuroscience behind Decision Making. http://bigthink.com/experts-corner/decisions-are-emotional-not-logical-the-neuroscience-behind-decision-making (2016).
7. Damasio, A. R.: The feeling of what happens: Body and emotion in the making of consciousness. Harcourt Brace, New York (1999).
8. Damasio, A. R.: Self Comes to Mind: Constructing the Conscious Brain. Pantheon, New York (2010).
9. Damasio, A. R.: Descartes’ Error: Emotion, Reason, and the Human Brain. Penguin, New York (1994).
10. Tranel, D.: Acquired sociopathy: the development of sociopathic behavior following focal brain damage. Progress in Experimental Personality & Psychopathology Research, 285-311 (1994).
11. Mendez, M. F., Chen, A. K., Shapira, J. S., & Miller, B. L.: Acquired Sociopathy and Frontotemporal Dementia. Dementia and Geriatric Cognitive Disorders 20, 99-104 (2005).
12. Minsky, M. L.: The Society of Mind. Simon and Schuster, New York (1986).

13. Dennett, D. C.: Why you can't make a computer that feels pain. Synthese 38 (3), 415-449 (1978).
14. Arbib, M. A., Fellous, J.-M.: Emotions: from brain to robot. TRENDS in Cognitive Sciences, 8(12), 554-561 (2004).
15. Balduzzi, D., Tononi, G.: Qualia: The Geometry of Integrated Information. PLOS Computational Biology. doi:http://dx.doi.org/10.1371/journal.pcbi.1000462 (2009)
16. Sellers, M.: Toward a comprehensive theory of emotion. *Biologically Inspired Cognitive Architectures 4*, 3-26 (2013).
17. Gomila, A., Amengual, A.: Moral emotions for autonomous agents. In J. Vallverdu, & D. Casacuberta, Handbook of research on synthetic emotions and sociable robotics (pp. 166-180). IGI Global, Hershey (2009).
18. McAuliffe, K.: This Is Your Brain On Parasites: How Tiny Creatures Manipulate Our Behavior and Shape Society. Houghton Mifflin Harcourt Publishing Co., New York (2016).
19. Kelley, D. J.: Modeling Emotions in a Computational System. http://transhumanity.net/modeling-emotions-in-a-computational-system (2016).
20. Maslow, A. H.: A Theory of Human Motivation. Psychological Review 50 (4) , 370-96 (1943).
21. Maslow, A. H.: Toward a psychology of being. D. Van Nostrand Company, New York (1968).
22. Ortony, A., Clore, G. L., Collins, A.: The Cognitive Struture of Emotions. Cambridge University Press, Cambridge (1988).
23. Bartneck, C., Lyons, M. J., Saerbeck, M.: The Relationship Between Emotion Models and Artificial Intelligence. Proceedings of the Workshop on The Role Of Emotion In Adaptive Behaviour&Cognitive Robotics http://www.bartneck.de/publications/2008/emotionAndAI/
24. Showers, A.: The Feelings Wheel Developed by Dr Gloria Willcox (2013). http://msaprilshowers.com/emotions/the-feelings-wheel-developed-by-dr-gloria-willcox
25. Plutchik, R.: The emotions: Facts, theories, and a new model. Random House, New York (1962).
26. Plutchik, R.: A general psychoevolutionary theory of emotion. In R. Plutchik, & H. Kellerman, Emotion: Theory, research, and experience: Vol. 1. Theories of emotion (pp. 3-33). Academic Publishers, New York (1980).
27. Plutchik, R.: Emotions and Life: Perspectives from Psychology, Biology, and Evolution. American Psychological Association, Washington DC (2002).
28. Norwood, G.: Emotions. http://www.deepermind.com/02clarty.htm (2011).
29. Hudak, S.: Emotional Cognitive Functions. In: Psychology, Personality & Emotion (2013). https://psychologyofemotion.wordpress.com/2013/12/27/emotional-cognitive-functions