Critical Nature of Emotions in Artificial General Intelligence

Key Nature of AGI Behavior and Behavioral Tuning in the Independent Core Observer Model Cognitive Architecture Based Systems

David J. Kelley[1,] Mark R. Waser[1, 2] and Arnold Sylvester[1]
[2]Digital Wisdom Institute, Vienna, VA; [1]Artificial General Intelligence Inc, Kent, WA
Mark.Waser@Wisdom.Digital, David@ArtificialGeneralIntelligenceInc.com

Abstract

This paper reviews the key factors driving the Independent Core Observer Model Cognitive Architecture for Artificial General Intelligence specific to modeling emotions used to drive motivational and decision making processes in humans; as it relates to or inspires the ICOM motivational systems. Emotions in ICOM are key elements of the ability to be self-motivating and make decisions. Behavioral tuning research case work around motivations in ICOM, as seen in the series 4 ICOM Isolation Studies designed to validate the series 4 model versus series 3 model and used to bench mark as well as tune the ICOM emotional processing core, are presented. Detailed is the reasoning for emotions in ICOM as used as a method of tagging ideas, concepts, and experiences for evaluation. Such emotions are the driving force behind the ICOM system's subjective experiences.

Keywords: Artificial Intelligence, Artificial General Intelligence, AGI, Self-Motivating, Cognitive Architecture, AI, AGI, ICOM, Independent Core Observer Model

1 Introduction

The Independent Core Observer Model (ICOM) research program was really developed out of a desire to focus on areas of research related to AGI [10] that had not been as widely researched as other areas. This eventually led to a focus on a system for artificial intelligence [9] that had the ability to have subjective experiences [25] and produce self-motivation driven by emotions. Such systems needed to be more robust then current self-motivating theory such as MicroPsi [21] to also include the internal emotional subjective experience along with retrospective behavior; where something like MicroPsi comes close in many ways however it falls short. ICOM in some ways is inspired by MicroPsi and builds on some key concepts in MicroPsi architecture. Designing such an emotion based system exactly like the human mind is not plausible given our current abilities; so instead of a bottom up approach to reverse engineering the human mind, which is better left to other teams, we approached the problem based on a logical approach to the effect of a system that has a subjective experience emotionally and was self-motivating and not dependent on any given substrate such as the biological human brain. In this case, I'll articulate the reason for emotions being key to the key concepts behind ICOM and how the ICOM system of emotional behavior is tuned in the most recent study completed by the team, being the Isolation Studies, creating the baseline for ICOM system behavior and conditioning.

Let us start with evidence that emotions are key to the human mind's ability to make decisions.

2 Humans, Emotions and our ability to make choices

Imagine a human that has a brain injury, through shoving a bit of rebar steel through his head but managed to survive with most of their intelligence intact. The one major issue is he loses his ability to feel emotions and further his inability to make decisions. This is in fact a true story from the 19th century. [1]

Another example is a patient with similar brain damage that had driven to the hospital on icy roads in terrible conditions; where he recounted his experiences en route logically and dispassionately, describing how he had avoided accidents by calmly applying the rules for driving on ice, while others about him were panicking and skidding and slamming on the brakes. Yet, when he had to decide between two dates for his next appointment, he spent half an hour listing the advantages and disadvantages for each of the proposed dates, until at last, in desperation he was told which date to come, whereupon the man thanked him, put away his diary, and left. [1]

Much of the research by neuroscientist Antonio Damasio documents many such cases in which the patient could not feel emotions and they lacked the ability to make decisions. [2] This brings us to the conclusions that, at least in humans, it appears that decisions are emotional driven and not logical per say at their root. Emotions are required to assign value [26] in the human mind therefore, if we are building a system that we want to behave like the human mind, we must include a structure like this that allows such ‘emotional’ value to be assigned. This conclusion led us to focus on emotional computational system capable of these subjective emotional states as a methodology for assigning value, making choices and being self-motivating.

3 Need for Emotions in Computing

M. Sellers noted in his paper on “Toward a Comprehensive Theory of Emotions for Biological and Artificial Agents” [3] that there is a need for emotions in computing and by extension Artificial Intelligence. To then build a system like this we designed a theoretical cognitive architecture [11] that would have an emotional model and decision making system at its core that would model logically the subjective experience, emotionally, of a human mind. Such systems are not without precedent as is the case in systems like MoNAD’s [17] or the aforementioned Seller’s [3] paper where the need is for emotions to help drive that self-evolving subjective experience.

For modeling that internal subjective experience we then turned to softer sciences, in particular clinical psychology, instead of systems as might be articulated by empirical psychologists which might use ANOVA (variance analysis), or factor analysis to model emotions that are focused on biology [4]. While these other models maybe be more measured in how they model elements of the biological implementation of emotions in the human mind at the biological level, we wanted to focus ICOM research solely from a logical standpoint so we turned towards the softer or clinical side of psychology were we looked at a number of systems. Our research is focused on 'how' of the logical modeling of those emotions or 'feelings'. The initial system we looked at was the Wilcox model [5] for human emotions but this turned out to be too cumbersome when there are systems like Plutchik [6] that are even simpler but still support a full range of human emotions.

It is ironic to note that Plutchik said that you couldn't model emotions mathematically; however, by reversing [7] the Plutchik model, it seems simple enough in the form of floating point decimal values where one value for each element of the Plutchik model can represent a given emotional state for each required valance and where combinations represent complex emotional states. In ICOM we decided that we would have a conscious and subconscious model representing that internal subjective emotional conditions and then ideas being represented as context trees or 'node' maps with emotional values associated with them based on various factors provided by the system. These node maps are essentially similar to hierarchical temporal memory [19] structures used to represent virtually everything in the ICOM architecture. ICOM also uses such trees in ideal modeling and as classification and regression trees [20] however that is outside the scope of this work.

Admittedly, this does get to a common problem called the Symbol Grounding problem [8] and while we might use 'names' to represent this item for human programmers, that 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. At least in ICOM, 'name' values don't mean anything but the emotional values are enforced by the system matrix and relationships and nothing more.

4 System Self Awareness

The Independent Core Observer Model architecture contends that consciousness [12] is a high level abstraction. And further, that consciousness is based on emotional context assignments evaluated based on other emotions related to the context of any given input or internal topic where the system is aware of this state in relationship to itself. These evaluations are related to needs and other emotional valences such as interests which are themselves emotions and used as the basis for 'value'; which drives interest and action which, in turn, creates the emergent [13] effect of a conscious mind. The largest complexity of the system is the abstracted subconscious and related system processing executing on the nuanced details of any physical action without the conscious mind dealing with direct details. Our ability to do this kind of decomposition is already approaching mastery in terms of the state of the art in technology to generate context from data [14]; or at least we are far enough along to know we have effectively solved the problem if not having it completely mastered it at this time.

A scientist studying the human mind suggests that consciousness is likely an emergent phenomenon. In other words, she is suggesting that, when we figure it out, we will likely find it to be an emergent quality of certain kinds of systems that occur under certain circumstances. [15] This particular system creates consciousness by design of the system.

It is really incorrect to say that consciousness in ICOM is an 'emergent phenomenon' in any way [16] even though colloquially we frequently use the term to explain ICOM but really, to be more precise, this is not the fact of the matter.

In ICOM consciousness is by design. The core engine itself is designed to processes the data in such a way as to produce awareness which is not even the most complex part of the system nor is it causing self-awareness really just be processing data but it's the way it does that emotional processing that is designed to interact with the rest of the system to create that abstracted awareness.

In future research ICOM will need to tackle subjective experience or ‘consciousness’ in a more controlled way that we can measure; where something like H. Porters work on Assessment of AI Consciousness [22] will need to be used to more objectively measure that work. Porters proposal itself is somewhat subjective and therefore will need to be baked out more to be less subjective so it can be used to measure consciousness in more detail without as much ‘subjectivity’.

Given this architecture, how do we get it to do something specific? We do this by biasing.

5 Application Biasing

ICOM is a general high level approach to overall artificial general intelligence. That being said, an AGI, by the fact that it is an ‘AGI’, should in theory be able to do any given task. Even before such a system has attained 'self-awareness', you should be able to train the system around certain tasks [17]. This is called biasing to a given application for usage of a specific ICOM instance.

By associating input or context with pleasure or other positive emotional stimulation, you can use those as a basis for the system to select certain actions. By limiting the action collection to what is possible in the given application and allowing the system to create and try various combinations of these actions, you essentially end up with an evolutionary algorithmic system for accomplishing tasks based on the amount of pleasure the system gains or, based on the system biases as might be currently present. Additionally, by conditioning, you can manipulate core context in ICOM to create a better emotional environment for the conditioning of a certain set of tasks you might want in an instance you are training.

In the training, or attempted biasing, keep in mind that personality or individual traits can develop in an implementation using ICOM if it’s allowed to store context as it is processed. Part of understanding that process in this research program has been something called the Isolation Study or studies which have been designed to allow us to see how the system behaves emotionally under the application of traumatic negative input.

6 The Isolation Study

The ICOM team has done 2 sets of isolation studies. We broke out the sets of experiments by the version of the core ICOM being used. At the time of this paper we are working with what we call the series 4 research [24]. The isolation study was first done with the series 3 [23] version of the core which used a simpler emotional model but focused on a build out around the Wilcox system and subsequently we moved to Plutchik as a basis for the core emotional models as we get the complexity of or the emotional landscape of the human mind as articulated by Plutchik without the computational overhead that the Wilcox models [5] produce. The full Plutchik version is the series 4 and we conducted the isolation study again with the series 4 as a comparative analysis.

As a side note, the series 4 version of the ICOM core is the first version able to have subjective experience and emotional landscape at the same complexity observed in humans.

In the series 3 isolation study we found that when the system was provided with input and after a time cut off and then provided input that it would perceive as 'pain' the system consciously reacted as expected but subconsciously seemed to enjoy the pain. After this series three experiment we actually spent 6 weeks working out the math by hand to-do the series 3 isolation study on paper to identify what was going on. In this original study, we ended up pulling in a lot of outside experts in psychology to help with that evaluation. The conclusion being that the system would rather have input then no input; so even if that input is negative, it is better than being cut off and alone based on how the ICOM model works.

The reason the isolation study is important is that it helps us tune the core matrix or learn how to bias the system to behave within human norms. Based on how ICOM works it can theoretically be unstable or mentally ill in terms of getting key emotional floating point values used in the matrix so far off that system responses are 'abnormal' by human standards. Given that key need to understand the effects of the various kinds of biasing in emotional responses, the isolation study currently is one of the best tools we have had for tuning the ICOM system.

In the series 4 Isolation study, which is relying entirely on a complex model based on Plutchik, we ended up running 3.15 million cycles or 'context' trees through the series 4 core conducting 40 separate cataloged experiment's which was split between 20 sets of control groups using random context trees and 10 sets of tests using random sets with isolation and 10 sets with sensory input (as opposed to random input) with isolation input. While the internal subjective experiences in ICOM are modeled with internal Plutchik model's the 2 sets of isolations study experiments compared to the control groups allowed us to see the system responded in like manor to the series 3 study but with a 'subjective' experience twice as complex.

7 Summary

The series 4 isolation studies showed that the series 4, while more complex did exhibit, what had appeared unusual results as seen in the series 3 work, but what should have been expected. This research helped tune the ICOM system to within human norm’s in terms of emotional evaluation of input. These kinds of experiments and understanding their implication and implementation effects on AGI systems is core to our ability to bias systems to preform or behave a certain way.

In this way, we reviewed the key factors driving the Independent Core Observer Model Cognitive Architecture specific to emotions used to drive motivational and decision making processes like humans as it relates to or inspires the ICOM motivational systems. In ICOM emotions are key elements of the ability to be self-motivating. We also reviewed case work around motivations in ICOM as seen in the series 4 ICOM Isolation Studies showing how the system emotionally behaved.

The series 4 isolation study sets the stage for additional research into tuning the working ICOM AGI Core to within human behavior norms in future studies such as the current human vs ICOM sentiment study using basic English.

References

[1] Damasio, Antonio *Descartes' Error: Emotion Reason and the Human Brain* Penguin Books 2005 ISBN: 014303622X

[2] Camp, Jim *Decisions Are Emotional not logical: The Neuroscience behind Decision Making* big think The Camp Negotiation Institute http://bigthink.com/experts-corner/decisions-are-emotional-not-logical-the-neuroscience-behind-decision-making

[3] Sellers, M *Toward a Comprehensive Theory of Emotion for Biological and Artificial Agents* Online Alchemy Inc., Austin Texas and Gotland University, Visby, Sweden 2013

[4] Milan, René email interview discussion on emotional modeling dated 10/10/2015

[5] Parrots, W.G. *Feelings Wheel Developed by Dr. Gloria Willcox* http://msaprilshowers.com/emotions/the-feelings-wheel-developed-by-dr-gloria-willcox/

[6] Norwood, G *The Plutchik Model of Emotions* Deeper Mind http://www.deepermind.com/02clarty.htm

[7] Lee, N. *Google It – Total Information Awareness Springer* http://www.springer.com/us/book/9781493964130 , http://www.amazon.com/Google-Information-Awareness-Newton-Lee/dp/1493964135/ ISBN 978-1-4939-6415-4

[8] Wikipedia Foundation *Symbol Grounding Problem* https://en.wikipedia.org/wiki/Symbol_grounding_problem

[9] Oxford Dictionaries *Artificial Intellignece* Oxford Dictionaries http://www.oxforddictionaries.com/us/definition/american_english/artificial-intelligence

[10] Goertzel B., *Artificial General Intelligence* doi:10:4249 Scholarpedia.31847 http://www.scholarpedia.org/article/Artificial_General_Intelligence

[11] ICT, *Cognitive Architecture* USC Institute for Creative Technologies http://cogarch.ict.usc.edu/

[12] Merriam-Webster's Learner's Dictionary *consciousness* Merriam-Webster http://www.merriam-webster.com/dictionary/consciousness

[13] Johnson J., Told, Andreas T, Sousa-Poza A., *A Theory of Emergence and Entropy in Systems of Systems* Elesevier, Procedia Computer Science Volume 20, 2013 – Complex Adaptive Systems http://www.sciencedirect.com/science/article/pii/S1877050913010740

[14] Malisiewicz, Tomasz *Deep Learning vs Machine Learning vs Pattern Recognition* vision.ai 2015 http://www.computervisionblog.com/2015/03/deep-learning-vs-machine-learning-vs.html

[15] Graham S., Weiner B., *Theories and Principles of Motivation* University of California from Cognition and Motivation http://www.unco.edu/cebs/psychology/kevinpugh/motivation_project/resources/graham_weiner96.pdf

[16] Yudkowsky, Eliezer *The Futility of Emergence* By Less Wrong Blog http://lesswrong.com/lw/iv/the_futility_of_emergence/

[17] CL Baldwin, Penaranda BN *Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification* US National Library of Medicine National Institutes of Heath http://www.ncbi.nlm.ni`h.gov/pubmed/21835243

[18] Sekiguchi R., Ebisawa H., and Takeno J. *Study on the Environmental Cognition of a Self-evolving Conscious System* 7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

[19] Hawkins J, Ahmad, S. *Properties of Sparse Distributed Representations and their Application to Hierarchical Temporal Memory* Cornell University Library

[20] Brownie, J *Classification and Regression Trees for Machine Learning* Machine Learning Algorithms http://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

[21] Bach J. *Modeling Motivation in MicoPsi 2* Massachusetts Institute of Technology, Cambridge, MA

[22] Porter H. *A Methodology for the Assessment of AI Consciousness* Portland State University, Portland Or cs.pdx.edu/~harry/musings/ConsciousnessAssessment.pdf

[23] Kelley, D. 2015 Series 3 Independent Core Observer Model Isolation Study Report Artificial General Intelligence Inc.

[24] Kelley, D. 2016 Series 4 Independent Core Observer Model Isolation Study Report Artificial General Intelligence Inc.

[25] Buck R., *What is this thing called Subjective Experience? Reflections on the Neuropsychology of Qualia* Neuropsychology, Vol 7(4), Oct 1993

[26] Maku, M *Physics of the Future – How Science will Shape Human Destiny and Our Daily Lifes by the Year 2100* by Random House Inc. 20111