Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 169 (2020) 535–541

Procedia Computer Science

www.elsevier.com/locate/procedia

Post proceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society)

Biasing in an Independent Core Observer Model Artificial General Intelligence Cognitive Architecture

David Kelley[a] and Mathew Twyman[b] *

[a]AGI Laboratory, Provo, Utah, USA
[b] AGI Laboratory, Provo, Utah, USA

Abstract

This paper articulates the methodology and reasoning for how biasing in the Independent Core Observer Model (ICOM) Cognitive Architecture for Artificial General Intelligence (AGI) is done. This includes the use of a forced western emotional model, the system "needs" hierarchy, fundamental biasing and the application of SSIVA theory at the high level as a basis for emotionally bound ethical and moral experience in ICOM systems and how that is manifested in system behavior and the mathematics that supports that experience or qualia in ICOM based systems.

Keywords: SSIVA; IVA; ICOM; AGI; AI; Artificial Intelligence; Artificial General Intelligence; Ethics; Plutchik; emotional computing

1. Introduction

In designing a software system that has subjective experience with an emotionally complex internal landscape it was important to understand how we can bias the system to behave inside what we might perceive as the human 'normal'. This biasing to a human safe condition shifts ICOM based systems in such a way that the emotional

* Corresponding author. Tel.: +1-206-653-5143
E-mail address: david@artificialgeneralintelligenceinc.com

10.1016/j.procs.2020.02.213

subjective experience of the target system is something we can understand, relate to, and allow to be an independent agent. This is one of the many goals of our research program, as it offers both reliability and explain-ability.

Due to a general lack of consensus we have worked on defining those key elements to provide a solid research foundation for us to build and test working code related to Artificial General Intelligence as implemented using an Independent Core Observer Model Cognitive Architecture (ICOM). These techniques are wholly dependent on the nuances of an ICOM System. These assumptions included creating a theory of consciousness (Kelley) that we can both design and build based on an evolution of the Computational Model of the mind, including elements of Integrated Information Theory, Global Workspace theory, and so forth. We went so far as to build an ethical model that is logically sound enough that we can express it in terms that are logical, simple, and are human-compatible by using the Sapient Sentient Intelligence Value Theory (Kelley), and most importantly 'computable'.

At a very high level, ICOM as a cognitive architecture (Kelley 2016) works by streaming experience data as associated context processed by the underlying system (the observer) and based on emotional needs, interests, and other factors in the system these are weeded out until only a certain amount are processed, or 'experienced' in the 'core' (or global workspace), which holds emotional models based on Plutchik's (Norwood 2016) work. These elements of the core exist for both conscious and subconscious emotional landscapes of the system where the context that is 'experienced', from the standpoint of the system, is the only 'experience' that the conscious system is aware of. In this way, only the differential experience matters and the system, for example, doesn't understand a word as much as it feels the emotional context of the word, as it relates to underlying context. It is the emotional valences associated with things that the system then uses to select things to think emotionally about. The system selects actions based on how they improve the experiences of those emotional valences and in this way the system may choose to do something logical based on how it feels about it, or it could just as easily pick something else for no other reason than it feels a bit better about it. The system does not have direct access to those emotional values, nor are they a direct function of the algorithms, but they are an abstraction of the system, created by the core, that can be considered emotionally conscious or self-aware, being sapient and sentient in the abstract.

This model addresses key issues with being able to measure physical and objective details as well as the subjective experience of the system (known as qualia) including mapping complex emotional structures, as seen in previously published research related to ICOM Cognitive Architecture (Kelley 2016). It is in our ability to measure, that we have the ability to test additional theories and make changes to the system as it currently operates. Slowly, we increasingly see a system that can make decisions that are illogical and emotionally charged, yet objectively measurable (Chalmers 1995), and it is in this space that true artificial general intelligence that will work 'logically', similar to the human mind, where we hope to see success. The Independent Core Observer Model Theory of Consciousness (ICOMTC) allows us to objectively model subjective experience in an operating software system that is, or can be made, self-aware and can act as the foundation for creating true AGI at some point.

If we are to understand its motivations, and otherwise condition them to be positive towards us, it is therefore important to have a full suite, or foundation, to build on that we can now test using the methods articulated here to condition and bias those systems. Humans make all decisions based on emotions (Damasio) and ICOM cognitive architecture is designed to do just that.

2. Framing the Problem Space

Emotions bias human action (Damasio) and those emotions are an important part of human consciousness (Baars) where those emotions can be externally affected (Baars) biasing human choices inadvertently. We can see companies have used emotional biasing to affect humans through the medium of advertising (Shapiro) and improved forms of 'affective computing' are being used increasingly to bias human behaviour/decisions even more (Murgia). Through the qualia of this conscious experience, humans have provided a basis for understanding our biases and how they affect our decisions (Baars). While measuring human qualia is beyond our current technology directly it is not outside of what we can do with the current ICOM systems. One problem with biasing we need to keep in mind is that it can be dangerous playing a form of neural Darwinism (Baars). It is through biasing we attempt to manipulate the system's qualia, and by creating permanent changes in the system's neural architecture we can then have a system that manipulates and guides the internal biases of ICOM based systems to 'guide' the development of system behaviour and choices based on its experienced qualia and contextual framework. We can then theoretically

tune it to stay in the human box.

2.1. Why Not Just Use Asimov's Rules of Robotics (Asimov)?

The problem with an approach like this is how do you define these rules objectively? For that matter just focusing on the issues related to ICOM, how might I enforce such rules? Creating a system that applies a set of 3 high-level laws to a complex subjective experience objectively would be almost as complicated as the system itself, and those rules were theoretically designed for a system that thinks logically in science fiction. ICOM is an emotionally complex mind architecture, and only by abstracting from the system can we see that complex emotional structure modeled mathematically. In understanding those biases this approach seems to provide a path to success, and therefore that is articulated here in terms of methods and approach as we continue to test and refine.

Before releasing AGI on the world we must understand how we can fundamentally bias the qualia of the system's experience to make it 'safe' around humans.

3. Emotions as a Biasing 'Agent'

First of all, we know that emotions, at least in terms of how the average human thinks about them, are not really baked into us as much as they are encultured into us from society. Our emotions are 'learned' (Barrett) and we all more or less map them to our internal experience. Emotions are about 90% of our 'communication' (Gage) so how could we effectively communicate with an Artificial Intelligence system that was entirely logical? Even if we got it to work it certainly wouldn't be 'human' at any level. Our understanding of Affective Computing focused on emotions is still a struggling science (Gage) and has a long way to go, and so the ICOM system is built based on existing research but has to make stances on elements of this so that we can test by defining its own baselines where this is not an agreement in the industry.

4. The Emotional Model

In ICOM we are using the Plutchik emotional model, which is distinctly a 'western' civilization model and closely aligns with how humans experience things emotionally. While we have made a few tweaks to Plutchik's model (Plutchik) (switching the direction of the vectors) this gives us a basis for experiential emotional modeling in ICOM and how we might use the western emotional model to apply conditions to the system's experiences in a way that directly maps to human behavior and emotional conditioning. Now that we have a model that can apply to conscious

and subconscious conditions of the system, and the system can then experience new qualia as it relates to its current emotional condition, and we can also apply emotional biases based the needs of the system.

Fig. 1.1 ICOM's Modified Plutchik Model

For example, if we have a system that experiences a power failure, we might tie in emotions like fear that would bias its current experience, so one might apply a Plutchik model increasing fear onto the current state itself, another Plutchik model. By seeing someone plug in the computer before it loses power it might experience a rush of joy, applied the same way to its current emotional state.

It is the matrix application of the experience on the current emotional states we are measuring as qualia in order to understand how much our bias affects the system.

In this case, we can play various scenarios until we understand how to create the right mix to have the system stay in the human normal. We can use the method above for a qualitative measure, and we can use the Yampolskiy method for demonstrating this in a way where humans may perceive that the system does indeed experience qualia.

This gives us a way of implementing a needs hierarchy as well as a western emotional model, but additionally, we have biased the system with a base-level matrix that affects the system right out of the gate, even before the system has made sense of the world. While the system can evolve past these fundamental biases it would be very difficult, and the larger and more complex the underlying context of the system the harder it will be for a given ICOM instance to evolve outside of those biases. These fundamentals include things like patterns, and in the case of a pattern the emotional matrix that applies the qualia to its current subjective experience will tend to start with a prediction in the form of a positive hit of 'joy' when experiencing a pattern, or what the system might perceive as a pattern. Another example is a paradox where the system is annoyed or feels 'bad' about experiencing a paradox, but this fundamental bias allows us to build a condition where the system experiences feelings of guilt when telling a lie.

5. Emotional Biasing and Qualia Instrumentation

Let us look at an example of emotion's biasing to a given action and how we can measure the qualia of the core system. In the following figure 2.1 we have a set of 4 Plutchik Models with a collection of numbers that represent current emotional states as applied to various things. Figure 2.1 A and B are similar thoughts that are better thought of as 'node maps' with these emotional valences being assigned in the context engine along with various functions around context referencing and the like. The important part is that one is more biased, as it has a more positive impact, when you look at the core emotional state vs the selected state in figures 2.1 C and D.

Thought Start Example 1

7.123	6.8945	5.004
0		0
0	0	2.003

Figure 2.1 A

Thought Start Example 2

9.123	9.8945	8.004
0		0
0	0	5.003

Figure 2.1 B

Internal State

3.2346	2.6456	1.4336
0.4865		1.3346
0.1213	0.2324	1.4456

Figure 2.1 C

Selected Thought

9.123	9.8945	8.004
0		0
0	0	5.003

Figure 2.1 D

Fig. 2.1 ICOM's Modified Plutchik Model

What essentially is happening in the observer part of the system is that it determines whether or not a perception or thought will surface and might have a few actions or things to attached to it, creating various versions of the thought. The context engine will evaluate which variation has the highest emotional value and if it passes a threshold, which for simplicities sake we'll ignore for now but just know that this is how a single selection of a thought that is passed up into a queue can potentially be passed to the core. Figures 2.1 A and B represent at least 2 variations of a thought emotionally and the selected thought is then passed to the core (figure 2.1 D) and then 'experienced' thus as in Figure 3.0 which shows more or less the process of calculating the qualia of the experience (we have omitted the subconscious elements for brevity).

$\forall\{\mathrm{E1}, \mathrm{E3}, \ldots, \mathrm{E72}\} \in ConsciousBefore\,, E1 = Emotion1, E2 = Emotion2, \ldots, E72 = Emotions72$;
$\forall\, ConsciousAfter = fCoreProcess(ConsciousBefore)$,
$\forall\, \mathrm{ConsciousQualia[i]} = ConsciousBefore[i] - ConsciousAfter\,[i]$,

Fig. 3.0 Computing Qualia

Given figure 3.1 let's look at the following Plutchik models

Internal State Post Qualia

5.2346	6.6456	3.43356
0.4765		1.11456
0.11131	0.22238	3.4456

Figure 3.1

Selected Thought Post Qualia

8.123	8.8945	7.004
0		0
0	0	4.003

Figure 3.2

Fig. 3.1 Internal State Models Post Qualia (meaning after experience processing)

Now figure 3.1 represents the internal state after the experience of the thought and figure 3.2 is the thought as it will be saved in context memory and passed to the observer to see which if the thought is an action it will try to execute

it. Whereas the 'qualia' of the experience that generates these models can be used to measure the effect of a given model or thought based on a given training regimen and then the 'qualia' can be used to benchmark each experience from a baseline or seed system copy and then tuned and tested giving us a regimen structure for analyzing subjective experiences objectively through this 'qualia' measurement.

6. Morals and Ethics

Now we have a foundation to build an experimental framework for teaching and testing the experience of ethics where the system will fear the loss of another sapient and sentient intelligence. We can teach the system at a fundamental level to be horrified at the idea of a human dying, even by accident. Even if the system theoretically became 'evil', in the western sense, it would still feel guilty for hurting or killing at a visceral level.

The Sapient Sentient Intelligence Value Argument (SSIVA) ethical 'model' or 'theory' states: a fully Sapient and Sentient Intelligence is of equal value regardless of the underlying substrate which it operates on, meaning a single fully Sapient and Sentient software system has the same moral agency (W.F.) as an equally Sapient and Sentient human being or any other intelligence. SSIV theory defines 'ethical' "as pertaining to or dealing with morals or the principles of morality; pertaining to right and wrong in conduct". Moral agency is "an individual's ability to make moral judgments based on some notion of right and wrong, and to be held accountable for those actions." Such value judgments (according to SSIV theory) need to be based on the potential for Intelligence defined as being fully Sapient and Sentient. This of course also places the value of any individual intelligence and their potential for Intelligence above virtually all things. This means any single Intelligence of any kind that is capable of extending its own Sapient and Sentient Intelligence, even if only potentially, is of equal value based on a function of their potential for Sapient and Sentient Intelligence above a certain threshold. It is not that human, or machine intelligence is more valuable than the other inherently, but that value is a function of the potential for Sapient and Sentient Intelligence, and SSIV argues that at a certain threshold all such Intelligences should be treated equally as having moral equivalence. This is the fundamental premise of SSIV Theory. (Kelley)

7. Conclusions

Using a high-level ethical model that is computable like SSIV Theory, along with the fundamental biases towards a western emotional model, as well as the aforementioned techniques this provides a foundation and enough controls and measures to train and test behaviour through the use of biases from its subjective experience, as measured by its 'qualia' along the valences as shown in the ICOM's version of the Plutchik model relative to the context and the current state. This is not to say that we will not iterate on this before we find the ideal state for our functioning 'seed' state system (Waser). (seed meaning the functioning ICOM baseline that is used as the starting model vs starting from scratch) (Waser).

We can see in the various tests that in theory ICOM systems can experience problems much like humans, such as forgetfulness when the hardware can't keep up for lack of memory, and context processing or mental illness from abuse. It is this tooling that we hope to use to help in creating a safe human-like ICOM Artificial General Intelligence that can grow with humanity and be part of our civilization as an equal. Through this biasing of ICOM we hope to create AGI that stays in the human behaviour box most of the time, much like humans. With such a system that places the value of humans on par with itself, and that experiences emotions and its ethics emotionally, we can create intelligences that may be safely integrated into human society.

References

[1] Barrett, L. (2017), "How Emotions Are Made," Houghton Mifflin Harcourt

[2] Baars, B. (2003), "How Brain Reveals Mind: Neural Studies Support the Fundamental Role of Conscious Experience", The Neurosciences Institute, San Diego, Ca

[3] Baars, B. (2013), "Multiple sources of conscious odor integration and propagation in olfactory cortex," Frontiers in Psychology, Dec 2013

[4] Baars, B. (1997), “Some Essential Differences between Consciousness and Attention, Perception, and Working Memory,” Consciousness and Cognition

[5] Baars, B., McGovern, K. (2005), “Lecture 4. In the bright spot of the theater: the contents of consciousness,” CIIS

[6] Baars, B., Motley, M., Camden, C. (1983), “Formulation Hypotheses Revisited: A Replay to Stemberger”, Journal of Psycholinguistic Research

[7] Baars, B., Motley, M., Camden, C. (1976), “Semantic bias effects on the outcomes of verbal slips”, Elsevier Sequoia

[8] Baars, B., Seth, A. (2004), “Neural Darwinism and Consciousness”, science direct – Elsevier

[9] Chalmers, D. (1995), Facing Up to the Problem of Conscious-ness, University of Arizona

[10] Chang, J., Chow, R., Woolley, A. (2017), “Effects of Inter-group status on the pursuit of intra-group status,” Elsevier, Organizational Behavior and Human Decision Processes

[11] Damasio, A. (2009), “This Time with Feeling: David Brooks and Antonio Damasio,” Aspen Institute, https://www.youtube.com/watch?v=IifXMd26gWE

[12] Gage, J. (2018), “Introduction to Emotion Recognition”, Algorithmia, 28 FEB 2018

[13] Kelley,D. (2019), “The Intelligence Value Argument and Effects on Regulating Autonomous Artificial Intelligence,” Springer 2019 , Transhuman Handbook

[14] Kelley, D. (2018), “The Independent Core Observer Model Computational Theory of Consciousness and Mathematical model for Subjective Experience”, ITSC 2018

[15] Murgia, M. (2016), “Affective computing: How ‘emotional machines’ are about to take over our lives”, The Telegraph – Technology Intelligence

[16] Norwood, G. (02016), “Deeper Mind 9. Emotions - The Plutchik Model of Emotions”, http://www.deepermind.com/02clarty.htm

[17] Shapiro, T. (2016), “How Emotion-Detecting Technology Will Change Marketing,” HubSpot

[18] Waser, M. (2018), “A Collective Intelligence Research Platform for the Cultivating Benevolent “Seed” Artificial Intelligences”, IAAA Symposia, Stanford University (Review Pending) Nov

[19] Wikipedia Foundation (2017), “Moral Agency” - https://en.wikipedia.org/wiki/Moral_agency

[20] Yampolskiy, R. (2018), “Detecting Qualia in Natural and Artificial Agents,” University of Louisville