Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 123 (2018) 221–227

Procedia Computer Science

www.elsevier.com/locate/procedia

8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017

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 fo-cused 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 implemen-tation of the simpler Plutchik model. We have seen responses that, at first glance, were surprising and seemingly illogical – but which mirror human re-sponses 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 on-ly known operational motivational system.

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures

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 envi-ronments – 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 hu-man-like cognition in a software system sufficiently

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures

10.1016/j.procs.2018.01.034

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 un-tested, and arguably dangerous, top-down hierarchical goal structure as their sole motivational driver, ICOM was specifically designed to have a human-like "emo-tional" 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 feel-ings, 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 de-mentia [11]. Clearly, it would be more consistent with human intelligence if our machine intelligences were implemented in the relatively well-understood cogni-tive state space of an emotional self rather than an unexplored one like unemotion-al and selfless "rationality".

While some might scoff at machines feeling pain or emotions or being con-scious, Minsky [12] was clear in his opinion that "The question is not whether intelligent machines can have any emotions, but whether machines can be intelli-gent 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 hu-man-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 evalu-ated or 'comprehended' and ready to decide how 'it' (being an ICOM implementa-tion) 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 prima-ry/current/conscious and a secondary/subconscious emotional state -- each repre-sented 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 emotion-al models. While the OCC model [22] has seemingly established itself as the standard model for machine emotion synthesis, it has the demonstrated [23] short-coming of requiring intelligence before emotion becomes possible. Since the Willcox "Feelings Wheel" [24] seemed the most sophisticated and 'logical' emo-tion-first model, we started with that. Unfortunately, its 72 categories ultimately proved to be over-complex and descriptive rather than generative.

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.