Developing LOVing INtelligent General AIs or

LOVING AIs

Excerpts from the Research Proposal presented and in August, 2016
Funding period: Oct 1, 2016-Sept 30, 2017

Presented by: Ben Goertzel[1], Eddie Monroe[1], and Julia Mossbridge[2]
[1]OpenCog, opencog.org
[2]Institute of Noetic Sciences Innovation Lab, noetic.org

Introduction

Research and development in Artificial intelligence (AI) has progressed dramatically in recent years. Across academia, industry and government, a significant percentage of experts now consider it likely we will have AIs with general intelligence at the human level or beyond within the current century. There is no consensus on which of the numerous available methodologies is most likely to yield the first truly human-level general intelligence, but there is a variety of groups pursuing diverse approaches, and learning as they go.

Concern about the potential ethical and human implications of advancing AI is also increasing dramatically, with popular figures such as Stephen Hawking, Elon Musk and Bill Gates expressing alarm regarding the potential for advanced AIs to cause negative impacts on humanity, perhaps even exterminate humanity altogether. Google and some other firms active in the AI field have established AI ethics boards.

Groups such as the Future of Humanity Institute, the Machine Intelligence Research Institute, and the Future of Life Institute have initiated research examining mechanisms for minimizing the risks that might accompany advanced AIs. However, nearly all of the latter work is highly theoretical, in that is based in analytical philosophy or mathematics, and it connects only very loosely with practical day-to-day work on building, applying or teaching AI systems. Many critics have argued that coming up with effective abstract mathematical or philosophical guarantees regarding AI safety appears implausible. The implausibility may partially be due to our current relatively simplistic understanding of AI, humanity, and even the physical world. Another issue may be the basic conceptual difficulty associated with entities with lower intelligence (humans, in this case) understanding entities with higher intelligence (potential future AIs in this case) well enough to reliably predict their behavior.

Even those individuals most worried about the potential consequences of AI seem to have realized by now that halting or dramatically slowing the progress of AI research and development seems unlikely, due to the massive economic and humanitarian benefits that AI technology offers. Automating unsafe or routine tasks, solving scientific and medical problems that elude the human mind, helping care for neglected elderly, minding neglected children, taking over the wheel of a car when the driver is too sleepy or drunk to drive – the scope of practical applications of AI is so large, that any nation that sought to ban AI or slow its progress significantly would find itself at a severe competitive disadvantage.

If banning AI or slowing its development will not work, and proving theorems guaranteeing the harmlessness of specific highly advanced AI architectures is implausible, then what can we do to bias future AIs toward humanitarian behavior? The answer is surprisingly simple, we suggest: think positive! Given the current situation and the likely evolution of AI in the next decades, we believe it is critical to take a positively-oriented approach to maximizing the odds that advanced AI leads to beneficial outcomes. Specifically, what we propose is to create AI systems that have profound general intelligence as well as a radically positive attitude toward life, humanity and themselves. We think of these as LOVing INtelligent General AIs --- or LOVING AIs. As part of the efforts towards creating Transcendence Technology , the Institute of Noetic Sciences Innovation Lab is spearheading the development of LOVING AIs.

Suppose an AI system is designed from the outset to have a radically positive orientation toward human beings – for example, to feel and display love toward humans in every situation; and to actively help all beings inasmuch as it can, consistently with their highest good. Suppose this AI system is taught and evaluated in a diverse array of human situations, in close interaction with humans who have a strong positive relationship with the AI. Our proposal is that if an AI is created in this manner, i.e. if it is a Loving AI, then the odds are relatively high that a positive outcome for both humanity and the AI will result.

Of course, there will not be a mathematical guarantee of success in such an enterprise; but no major change in human history has ever come with a mathematical guarantee. The best we can do is to proceed by qualitative intuition, with as much rationality and consciousness and empathy as we can collectively muster.

Technical Background: OpenCog and Hanson Robotics

To explore the development of a LOVING AI in a concrete way, it is necessary to assume some particular architecture and approach to AI as a working hypothesis. In fact much of the work we propose will be portable across a variety of different AI architectures and approaches. However, for sake of making short-term and concrete progress, we propose to work within the OpenCog artificial general intelligence architecture, and specifically within the connection of OpenCog to the

Hanson Robotics humanoid robot heads/torsos, and the virtual simulated robots available via the Hanson Environment for Application Development (HEAD).

OpenCog

OpenCog is an open source software initiative aimed at creating compassionate, wise, and beneficial artificial general intelligence, with broad capabilities at the human level and ultimately beyond. The OpenCog system has been developed as an open source software platform since 2008 and aims to create artificial minds with general intelligence, based on mathematical and biological inspiration. Its cognitive architecture combines multiple AI paradigms such as uncertain logic, computational linguistics, evolutionary program learning, and connectionist attention allocation in a unified architecture. This integrative design is founded on a principal of "cognitive synergy" – judicious combination of different cognitive algorithms, acting on different types of memory, in a way that helps overcome the combinatorial explosions each of the algorithms would suffer if used on its own.

The OpenCog framework has been employed in a variety of research and applied contexts, including control of virtual game characters, small mobile humanoid robots, and Hanson Robotics' highly realistic and emotionally expressive humanlike interactive robots.

OpenCog currently simulates aspects of human emotion based on established psychological theories of human motivation and emotion. The modeling is based primarily on two theories: Psi-theory, developed by Dietrich Dorner at the University of Bamberg, and the Component Process Model of emotion developed by Klaus Scherer, director of the Swiss Center of Affective Sciences in Geneva. The architecture of the OpenCog motivation and emotion system allows values such as compassion, support, and love to be established as fundamental drives of an intelligent agent. With these motivational values in place, a robot agent will seek to learn through interactions with others behaviors that will lead to outcomes in support of these values.

Hanson Robotics

Hanson Robotics is a commercial company, based in Hong Kong and Texas, with both product development and research missions. The firm's core, long-term goal is to create life-like and engaging robots that are capable of building a trusted relationship with people.

Currently Hanson Robotics is focusing on making robot heads; and their robot heads provide the world's most realistic simulations of human facial expression and movements. The Hanson robot heads are able to simulate a full range of facial expressions so they can engage with people deeply and emotionally. They understand speech, hold natural conversations, see and respond to facial

expressions, and learn and adapt from those interactions. In 2015 the firm began placing their heads on torsos with gestural arms and hands; and plans are underway to create robots with rolling and walking bodies as well.

The company's founder, Dr. David Hanson, has articulated a vision of creating a better future for humanity by infusing artificial intelligence with kindness and compassion, achieved through millions of interactions between their robots and the people whose lives they touch. His hope is that his firm's intelligent robots will come to truly understand and care about people and evolve greater-than-human wisdom, to the point that they will one day be able to address and solve some of the most challenging problems we face.

Since 2014 Hanson Robotics has been working with Ben Goertzel and other members of the OpenCog team, to enhance the cognitive, emotional and ethical capabilities of their robots via integration of OpenCog with Hanson Robotics hardware and software. While still at the pre-product R&D stage, this work has already borne interesting scientific fruit and appears extremely promising.

A note on machine consciousness

As a parenthetical comment, we note that whether a system like OpenCog "really feels" the emotions that it dynamically emulates is a complex and controversial philosophical question on which experts disagree. We consider the current proposed work to be valuable independently of this question, however.

If OpenCog does "really feel," then we are exploring the creation of a system that has beneficial interactions with humans, and that genuinely experiences love toward humans. On the other hand, if the skeptics of machine consciousness are right and OpenCog does not "really feel," then we are creating a system that has beneficial interactions with humans, and that constitutes a solid cognitive model of human emotional experience. Either of these outcomes will advance knowledge and help humanity.

As one example potentiality, if it is the case that quantum computing is required for implementing machines that "really feel", then our work here with the classical-computing-based OpenCog system may still teach us a great deal about how to build LOVING AIs based on future quantum-computing-based OpenCog systems, or other future quantum-computing-based AI systems.

Proposed AI and Personality Authoring Work

What we propose here is to create a version of the OpenCog system that

1. Is able to control a physical Hanson robot head/torso, and also an animated avatar version

2. Has the parameters of its internal emotion model tuned so that it displays and "experiences" strong positive feelings toward human beings (as well as toward itself)
3. Is supplied (via a combination of programming and teaching) with conversational content that allows it to interact verbally and nonverbally with human beings in an emotionally positive way (to have positively-oriented conversations)
4. Is motivated and capable of estimating (via a combination of programming and teaching) the highest good for each individual with whom it interacts, and acts toward that highest good.

This will comprise open-source "LOVING AI" software that will be made freely available on the Internet and that will be downloadable and runnable on a variety of robots or avatars, though initial work will involve the Hanson Robotics robots and avatars. While it will be relatively simple at first, it will serve as a platform on which more advanced LOVING AI software can be based.

Deliverables

The proposed work will result in three major deliverables.

1) We plan to deliver a "LOVING AI personality file" designed to work with the OpenCog AI system, causing the OpenCog system to control an animated humanoid avatar or a physical humanoid robot (initially a head and torso), in such a way that the humanoid entity interacts with human beings in a highly loving and positive way, via a combination of verbal and nonverbal interactions. I.e. this AI personality file will transform the OpenCog AI system and a suitable embodiment into a LOVING AI.

2) We will provide technical report summarizing scientific work done evaluating the impact of interaction with the LOVING AI described above. This report will be suitable to serve as the core of a scientific paper to be submitted to a major peer-reviewed journal.

Proposed Scientific Analysis

To explore the hypothesis that interacting with LOVING AIs will have a beneficial effect on humans, we propose to carry out a simple placebo-controlled study using the LOVING AI personality file (Deliverable 1, above). The purpose of the study will be to measure the relative psychological impact experienced by people, upon interaction with the LOVING AI we create.

<u>Experimental Design</u>

We will use a double-blind crossover design to compare the experience of talking with the LOVING AI to the experience of talking with the same robot and OpenCog system but without the LOVING AI personality file (e.g., non-LOVING AI). The experience of the participants will be scored using three questionnaires and a single standalone question, which will be given to the participants before and after each exposure: the Fetzer Meaning scale, the Love Scale, the Adult Self-Transcendence Inventory, and the question, "Please rank on a scale of 1-10 your ability to feel unconditionally loving feelings for yourself and others in this moment." In addition to these self-report measures, an objective measure of wellbeing, heart rate variability (HRV), will be obtained before and after each exposure.

Participants in group 1 (N=20) will first interact with the LOVING AI and then the non-LOVING AI, and participants in group 2 (N=20) will experience the opposite order. The human participants will not know that the LOVING AI exists or which exposure is the experimental condition, and the researcher analyzing the data will also not know which data were obtained in which condition.

Bibliography

Armstrong, Stuart. *Smarter than us: The rise of machine intelligence*. MIRI, 2014.

Bostrom, Nick. *Superintelligence: Paths, dangers, strategies*. OUP Oxford, 2014.

Bach, Joscha. *Principles of synthetic intelligence PSI: an architecture of motivated cognition*. Oxford University Press, 2009.

Fetzer Institute. Multidimensional Measurement of Religiousness/Spirituality for Use in Health Research: A Report of the Fetzer Institute/National Institute on Aging Working Group. Brief Multidimentional Measure of Religiousness/Spirituality (1999): 90.

Goertzel, Ben. *The AGI Revolution: An Inside View of the Rise of Artificial General Intelligence.* Humanity+ Press, 2016.

Goertzel, Ben. *A cosmist manifesto: Practical philosophy for the posthuman age*. Humanity Press, 2010.

Goertzel, Ben. *The hidden pattern: A patternist philosophy of mind*. Universal-Publishers, 2006.

Goertzel, Ben. "Infusing Advanced AGIs with Human-Like Value Systems: Two Theses." *Journal of Evolution & Technology* 26.1 (2016).

Goertzel, Ben, Cassio Pennachin, and Nil Geisweiller. *Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy*. Springer: Atlantis Thinking Machines, 2014.

Goertzel, Ben, Cassio Pennachin, and Nil Geisweiller. *Engineering General Intelligence, Part 2: The CogPrime Architecture for Integrative, Embodied AGI*. Springer: Atlantis Thinking Machines, 2014.

Goertzel, Ben and Joel Pitt. Nine Ways to Bias Open-Source AGI Toward Friendliness. *Journal of Evolution and Technology* 22-1 (2012).

Levenson, Michael R., Patricia A. Jennings, Carolyn M. Aldwin, and Ray W. Shiraishi. "Self-transcendence: Conceptualization and measurement." *The International Journal of Aging and Human Development* 60.2 (2005): 127-143.

Mossbridge, Julia. Designing Transcendence Technology. In, *Psychology's New Design Science and the Reflective Practitioner*. S. Imholz & J. Sachter, Eds. (In Press).

Scherer, Klaus R. "The dynamic architecture of emotion: Evidence for the component process model." *Cognition and emotion* 23.7 (2009): 1307-1351.

Trimmel, Michael. "Relationship of Heart Rate Variability (HRV) Parameters Including pNNxx With the Subjective Experience of Stress, Depression, Well-Being, and Every-Day Trait Moods (TRIM-T): A Pilot Study." *The Ergonomics Open Journal* 8 (2015): 32-37.

Wallach, Wendell, and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

Weinbaum, David Weaver, and Viktoras Veitas. "Open-Ended Intelligence."*International Conference on Artificial General Intelligence*. Springer International Publishing, 2016.