AGI and Neuroscience: Open Sourcing the Brain

Randal A. Koene[1]

[1] Halcyon Molecular, Carboncopies, 505 Penobscot Dr.,
Redwood City, CA 94063
r@halcyonmolecular.com, Randal.A.Koene@carboncopies.org

Abstract. Can research into artificial general intelligence actually benefit from research in neuroscience and vice-versa? Many AGI researchers are interested in how the human mind works. After all, within reasonable limits we can posit that the human mind is a working general intelligence. There is also a strong connection between work on human enhancement and AGI. We note though, that there are serious limitations to the use of cognitive models as inspiration for the components deemed necessary to produce general intelligence. By delving deeper into the neuroscience, missing and hidden functions and global interactions can be uncovered. In particular, this is possible by explicitly uncovering the map of brain circuitry at a scope and a resolution that is required to emulate brain functions.

Keywords: Artificial intelligence, neuroscience, human mind, general intelligence, hidden functions, brain emulation, substrate-independent minds, human enhancement.

1 Introduction

For the past few years I have been keenly interested in and involved with the growing community of researchers in artificial general intelligence (AGI), even though I am by training a computational neuroscientist and, through *carboncopies.org*, actively working on the implementation of functions of mind that are based explicitly on the architecture of biological brains. I have participated in a number of AGI-related events, such as the AGI conferences of 2008 and 2010. I share with some of the pioneers of AGI (e.g. Ben Goertzel [1]) the conviction that there are areas of overlap between AGI research and neuroscience research, in which an interdisciplinary perspective is of particular value.

Still, there have been recurring questions, asking whether such mutual benefit truly exists. To my knowledge, those questions have not yet been addressed concretely in front of gathered experts of both fields of research. Are investigations about biological brains that cross boundaries of scale and resolution, such as the Blue Brain project [2] going to lead to understanding of the essentials of general intelligence? Or will the mathematical study of optimal universal artificial intelligence [3] lead to

actual implementations of AGI? In this position statement, I outline the manner in which I intend to address the relationship between AGI and neuroscience.

1.1 Perspective

Let us take a step back to gain some perspective. It is worthwhile to consider why we are interested in strong AI or AGI. Pei Wang notes that "[of course the goal of AI research is] to make computers that are similar to the human mind"[4]. We are reminded of speculative scenes with robotic assistants and play-pals, brought to us by the enthusiastic projections depicted in 1960s television extravaganzas. How much of AGI is about creating new thinking beings that are as versatile as we are, so that they can fulfill a multitude of the same roles that we can? How much of it is a matter of looking for the machine solution to those things that we cannot do, those things perhaps, which we wish we could do?

For example, it is very difficult for us to work at a complex job around the clock for many days without interruption; many of us wish that we could. It is difficult for us to take care of a multitude of tasks concurrently; we wish we could. Delegating the tasks to intelligent machines may seem preferable to delegating them to other human beings. There are also some mental tasks that are not a good match to the design of our minds, and yet even tasks that to us seem obviously related may represent a pool of requirements so general that adaptation is needed in order to tackle each new problem.

I suspect that we would like to be able to do all of this ourselves if we could. There is an analogous situation with regard to physical tasks. To fly we design airplanes. To hear the echolocation of bats we design sensitive microphones. Yet the superheros of our imagination are those individuals who incorporate those capabilities within themselves. We wish that we could carry out all of the perceptual and mental operations as well, because then we would grow to have new sensations and the ability to understand and experience that which is at present beyond us. There we have it, a clear connection between the search for human enhancement and the drives that motivate work in AGI.

2 AGI and the Human Brain

Some AGI researchers are explicitly pursuing forms of (general) intelligence designed from first principles and without a desire for comparability or compatibility with human intelligence. By and large though, many of the underlying objectives that drive the search for AGI also involve an interest in anthropomorphic interpretations of intelligent behavior. The operations of the human mind are of interest to strong AI and many in the field of AGI [1,4].

Abstract, optimized intelligence is certainly of theoretical interest [3], but if it is not feasible then it will have limited practical impact on progress in research and

development [5]. Speaking of practical endeavors, how does AGI earn its capital letters? All AGI projects aim to produce something that has not existed before it was explicitly created, making it artificial. AGI projects aim to produce an implementation that is able to carry out problem solving in at least one domain, usually with the ability to learn and adapt its problem solving approach. These are some of the hallmarks of intelligence [6,7].

The special focus in AGI is on the aim to create something with more general applicability, able to apply its intelligent processes to a variety of problems. By analogy with human intelligence, imagine the scenario where we have the ability to read, as well as knowledge of mathematics. We can then read textbooks and articles in order to work in related disciplines, such as computational neuroscience or macroeconomics. According to such criteria for generality, the human mind is an example of a general intelligence, even if it is not a universal general intelligence [3].

2.1 High-level Insight from Psychology and Cognitive Science

In past decades, research in AI has been guided by insights about the human mind from experimental and theoretical work in psychology and cognitive science. Insights at that level were the obvious source of information, since very little was known about the underlying mechanistic architecture and functionality of the brain.

Characteristics of the cognitive architecture of the human mind, modularity and functional specialization, such as expressed in ACT-R [8], SOAR [9,10], reinforcement learning [11], cognitive models of the hierarchical visual system [12], etc., can be derived through experimental procedures such as psychophysics, through introspection, and through some verification by neuroscientific experiments (e.g. neuroscience carried out in the visual system [13]).

For a long time it has been impossible in neuroscience to reconcile the very small with the very large. Investigation at large scale and low resolution was congruent with cognitive science, and led to the identification of centers of the brain responsible for different cognitive tasks through fMRI studies [e.g. 14]. Having a rough map of the localization of the known set of gross cognitive functions within a standardized brain does not actually add significantly to an understanding of exactly how the brain does what it does.

By contrast, psychophysical experiments can be used to determine parameters, limits, error modes. This sorts out some of the ways in which the mind's functions do work and some of the ways in which they do not. That data sheds some light on underlying algorithms that we may infer [15].

The problem with this approach is that it can only illuminate the treatment of that feature of behavior which is being tested. Like all studies that are in effect variations of sensitivity analysis [16,17] of a black-box model, it can measure effects and enable reverse engineering of the I/O functions only for those cases that are expressed[1].

1 In formal sensitivity analysis, this is related to the known pitfalls of “piecewise sensitivity”, where analysis can take into consideration only one sub-model at a time. Interactions among factors in different sub-models may be overlooked, a so-called Type II error. In the case of

Hidden functions are not included, the massive and complex interactions between the neuronal groups involved in different modules of the brain are not unlocked, potentially overlooking critical aspects of collaborative operation.

Traditional neuroscience, on the other hand, which offers studies at resolutions greater than the behavioral and the cognitive, was limited to the careful examination of very specific aspects of brain physiology and dynamics. Those studies tells us something about the characteristics and intricacies of the substrate in which the functions that are carried out by the mind are embedded, but they cannot give us an understanding or a way to grasp those functions in an algorithmic manner.

2.2 Access to Insights from Neuroscience

The younger domains of computational neuroscience and neuroinformatics are finally beginning to produce results that close the gap between the "big-picture" abstractions and the physiological detail. The computational approaches are able to use functional models of components of the brain and to combine those with structural information from the "connectome" that explains how the components can interact [18]. Of course, these results are only just emerging. It is still difficult to validate models, to conclusively pick one more correct model among many. In any case, current models are constructs that are based largely on the consensus interpretation of observed characteristic structure and function in an inhomogeneous collection of samples.

As models in computational neuroscience do increasingly provide reliable insights, those directly address a main problem faced by the field of AGI. The hurdles in AGI are not that we cannot think of enough wonderful capabilities that the machine should have, but they are a dearth of knowledge about how to implement those capabilities. If we knew how to implement them, strong AI would be a reality now. The various researchers, labs and companies involved in the field are exploring their best conceptions of what a successful implementation might be. The brain's implementation is not necessarily the best one according to criteria used to measure performance at solving a particular problem, but at the least it is an existing implementation, and we have some idea of the specifications that it meets.

2.3 Should AGI Learn from the Human Brain?

An important thing that AGI can learn from the brain is how you integrate and coordinate modules of a complex system in such a way that the result is self-consistent, fairly robust and capable of some adaptation [19,20]. How do you self-organize associations at different levels of abstraction? How do you include the causal

---

the human mind, only a small subset of possible sub-models may be considered at all, which can lead to a so-called Type III error, by potentially analyzing the wrong problem.

dimension? How do you use the co-design of hardware implementations with their target functions as a way to produce an efficient autonomous system?

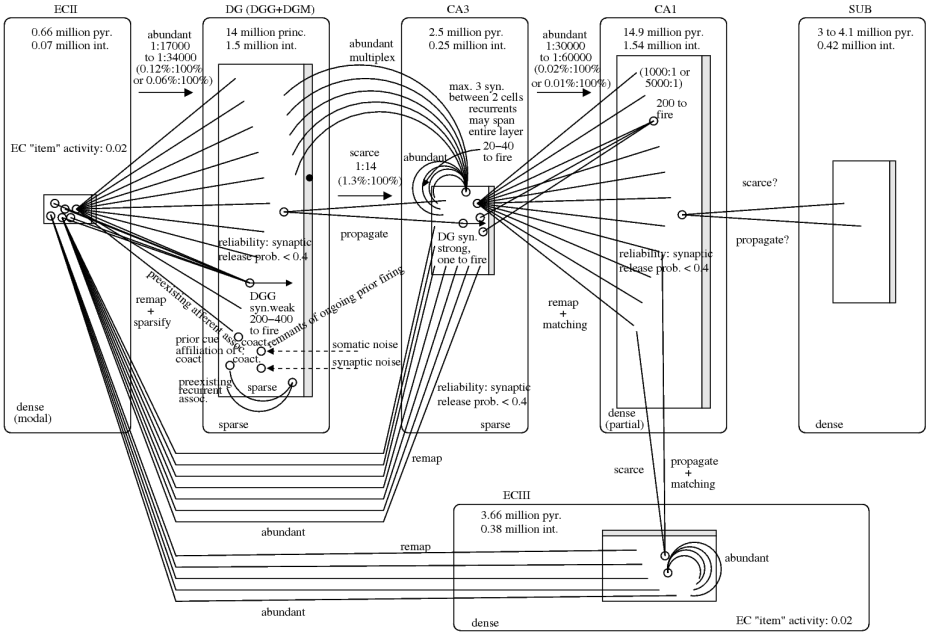

Fig.1. This figure depicts the conceptual considerations of a part of a computational simulation that is based on a neuroscience model of the acquisition of declarative memory and its eventual integration with procedural memory [21,22,23]. We note the involvement of different modules that employ different physical mechanisms, different forms of storage and representation, at different time-scales. By incorporating more details of the structure and the physiology, models such as these lend themselves to the discovery of interactions across modalities, between specific circuits and therefore of hidden functions.

A recurring argument against borrowing from neuroscience in the development of AGI has been to note that the low-level design of the brain is very complex, possibly needlessly complex for general intelligence [24]. In principle, this is a relevant argument. But other difficulties arise when one instead chooses the most obvious alternative approach: Observe high-level processes and implement those.

To confidently apply that method, the high-level observations need to capture the essential aspects of general intelligence. This is not obviously the case, for example, because that would require a-priori insight into the (ideally one-to-one) correlation between observed activity and abstract function. And how do we know when we have observed, in operation, all the relative functions? We must wonder if the brain's implementation details are indeed so much more complex than necessary. If they are, then why does our intelligence require a human brain? Why can equivalent intelligence not appear equally well in the brain of a chimpanzee, or a cat?

Let me use an analogy to succinctly raise my concerns about the strong reliance in AGI research on obviously vastly simplified models of cognition. If you were attempting to reverse engineer a CPU in order to discover all of the functions embedded in its microcircuitry, would you restrict yourself to the observation of five cherry-picked programs running on the CPU? Especially, would you do so if those five were picked, because they were the easiest ones to characterize, since none of the five happen to use a sequence of more than three distinct operations? How would this give you good insight into the essence of what makes that CPU good at its tasks? The aspects of cognition that are well-explained by the popular cognitive architectures cited in AGI research are similarly based, in part, on cherry-picked experiments and corresponding data about human cognitive processes [25].

Note that I am not suggesting that intelligence could not be implemented using different primitives, such as a rigorous set of algorithms. I am suggesting that learning from the brain's gross activity, but not its detailed interactions may be missing essential ingredients of general intelligence.

3 Brain Emulation as a Route to AGI

For many years, I have been deeply involved in efforts to reverse engineer, re-implement and emulate the operations of the brain that are essential for the dynamic functions of the mind. The prospects for truly large scale high resolution emulations of reimplemented brain are rapidly improving. Once this is achieved, it will be possible to run a mind on another substrate and to move the emulators and data between different substrates, effectively making mind functions substrate-independent.

So, from the cognitive science, you can identify functional capabilities that are of specific interest to a strong or general AI. In neuroscience, we investigate examples of the implementation of such functions. Learning from these implementations is akin to the way in which a programmer can learn by studying the code produced by others, which is one of the underpinnings of the open source movement. One of the great gains from brain emulation is access to the raw functions and parameters of the mind. Brain emulation open sources the implementation of the human mind.

3.1 Is an AGI a Substrate-Independent Mind? Is a SIM an Artificial General Intelligence?

If we understand the algorithms upon which an AGI is built, then it is possible to compute those algorithms in a variety of computing platforms. The AGI is effectively substrate-independent, and a sufficiently advanced AGI could therefore constitute a substrate-independent mind (SIM).

Conversely, if, as stated above, we accept that the definitions of the G and the I in AGI can apply to human minds, then a substrate-independent implementation of a human mind is an artificial version of the necessary functions. That makes the substrate-independent mind an AGI. Either perspective appears tenable. That there is a branch of AGI research that focuses explicitly on routes to SIM such as the relatively conservative implementation known as whole brain emulation (WBE) is immediately apparent from the Wikipedia entry on Strong AI and AGI [26].

3.2 Can we Produce a SIM without Understanding the Mind?

Theoretically, it is possible to create a substrate-independent mind without understanding how the functions of the mind work at all relevant levels of abstraction. For example, this could be achieved by a procedure that results in whole brain emulation at some acceptable resolution. It would be possible to identify the connectome and to identify each component and its intrinsic operation. Feasibility of this approach is as yet unproven. Without understanding more about the mind it is difficult to verify that data is acquired at the resolution that is necessary. It is also very difficult to test whether a function was correctly re-implemented. It is therefore not likely that a SIM would be created without any understanding of the mind. But it is also unlikely that a first SIM would require a total understanding of the mind at all scope and all resolution.

Emulation is a concrete approach. If carried out conscientiously, the readily apparent connection with an existing physical ground-truth offers some guarantees that such a method will be able to produce a general intelligence. Consequently, the goals for substrate-independent minds, as described with initial technology agnosticism at carboncopies.org by its founders, are also concrete ways to address AGI.

4 Concluding Remarks

Open sourcing the brain, learning directly from it, or from the reimplementation of some or all of its parts is the most potent contribution to a fruitful bi-directional exchange of knowledge between the fields of AI and neuroscience. I propose that there is a novel effort with actions to pursue here: We can discover if there are still elements of a whole brain that are essential to general intelligence, but that have so far been overlooked. We can determine if the requisite size and complexity of intelligent processing implies that hardware is still a hurdle. Does a feasible approach demand massive parallelism such as in neuromorphic hardware perhaps? And we may learn whether generality can be accomplished only through embodiment or total immersion in the context of a problem space, a realistic environment.

The process of laying bare the corpus and the elements of the brain in its full scope and at the necessary resolution depends on new tools, which are a topic ripe for

another occasion. New tools are inextricably implicated in the rise of new paradigms and in the occurrence of scientific revolutions. At the very least, using cutting-edge tools to open source the brain will bring many more creative minds to the task of reverse engineering the one working implementation of general intelligence.

Doing that will also bring us closer to that *wish* which I uncovered in the Introduction. We approach the ability to enhance our own mental capabilities and perceptions. When we arrive at that point we have to wonder: Would we rather that strong AI exists mostly in separation from us, or would we rather that the the same capabilities are extensions of ourselves? To borrow an argument [27]: "*How can AI be 'more than human' if it is something different entirely? Is an apple 'more than an orange'? One may taste better, and one may be juicer, but an apple is not an 'enhanced orange' nor is an orange an 'trans-apple'.*

If you could run a million different algorithms in parallel and carry out tasks all over the globe, being fully aware of them, but not bogged down by them, would you? Or would you wish to continue to inhabit the constrained perception that we have right now, leaving the grand network largely to *de novo* intelligences? It is a vast parameter space to explore and a challenge to optimize successfully, but pioneering experts will lead this field for enhancement as for novel AGI. If we can reverse engineer the brain sufficiently so that we can both learn from it and add to it, then perhaps we should put a new spin on Minsky's famous quote: *Will robots inherit the Earth? Yes, but they will be us.*

Acknowledgments. I thank Anders Sandberg, Demis Hassabis, Ben Goertzel, Suzanne Gildert and all my friends at Halcyon for numerous and deep conversations involving the relationship between AI and neuroscience, which planted the seeds for the arguments I present.

References

1. Goertzel, B., Pennachin, C.: Artificial General Intelligence. Springer, New York (2007)
2. Markram, H.: The Blue Brain Project. In: Nature Reviews Neuroscience, vol. 7, pp. 153--60 (2006)
3. Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin (2004)
4. Wang, P.: Artificial General Intelligence: A Gentle Introduction, http://sites.google.com/site/narswang/home/agi-introduction
5. Gildert, S.: Pavlov's AI: What do superintelligences REALLY want?. At: Humanity+ @Caltech, Pasadena, CA (2010)
6. Luger, G.F.: Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 6th Edition. Addison-Wesley, New York (2008)
7. Burns, N.R., Lee, M.D., Vickers, D.: Individual Differences in Problem Solving and Intelligence. In: Journal of Problem Solving (2006)
8. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. In: Psychological Review, pp.1036-60 (2004)

9. Laird, J., Newell, A., Rosenbloom, P.: SOAR: an architecture for general intelligence. In: Journal of Artificial Intelligence, vol. 33(1), pp.1-63 (1987)
10. Lehman, J.F., Laird, J., Rosenbloom, P.: A Gentle Introduction to SOAR: An Architecture for Human Cognition: 2006 Update. (2006)
11. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA (1998)
12. Marr, D., Ullman, S. Poggio, T.: Vision. A Computational Investigation into the Human Representation and Processing of Visual Information. MIT Press, Cambridge, MA (2010)
13. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. In: Journal of Physiology, vol. 160, pp.106-154 (1962)
14. Op de Beek, H.P., Haushofer, J., Kanwisher, N.G.: Interpreting fMRI data: maps, modules and dimensions. In: Nature Reviews Neuroscience, vol. 9, pp. 123-135 (2008)
15. Geissler, H.-G., Link, S.W., Townsend, J.T. (Eds.): Cognition, Information Processing, and Psychophysics: Basic Issues, Erlbaum, Hillsdale, NJ (1992)
16. Saltelli, A, Tarantola, S., Chan, K.: Quantitative model-independent method for global sensitivity analysis of model output. In: Technometrics, vol. 41(1), pp. 39-56 (1999)
17. Winsberg, E.: Simulations, models and theories: Complex physical systems and their representations. In: Philosophy of Science, vol. 68(3), Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers (Sep., 2001), pp. S442-S454 (2000)
18. Sporns, O., Tononi, G., Kötter, R.: The Human Connectome: A Structural Description of the Human Brain. In: PloS Computational Biology, vol. 1(4), e42 (2005)
19. Hassabis, D.: Combining systems neuroscience and machine learning: a new approach to AGI. At: The Singularity Summit '10, San Francisco, CA (2010)
20. Koene, R.A.: The 25 Watt bio-computer: Lessons for Artificial Human Intelligence and Substrate-Independent Minds. At: Humanity+ @Caltech, Pasadena, CA (2010)
21. Koene, R.A.: Functional requirements determine relevant ingredients to model for on-line acquisition of context dependent memory. Ph.D. Dissertation, McGill University, Montreal, Canada (2001)
22. Koene, R.A., Hasselmo, M.E.: First-in-first-out item replacement in a model of short-term memory based on persistent spiking. In: Cerebral Cortex, vol. 17(8), pp. 1766-81 (2007)
23. Koene, R.A., Hasselmo, M.E.: Reversed and forward buffering of behavioral spike sequences enables retrospective and prospective retrieval in hippocampal regions CA3 and CA1. In: Neural Networks, vol. 21(2-3), pp. 276-88 (2008)
24. Gorelik, D.: Reducing AGI complexity: copy only high level brain design, http://aidevelopment.blogspot.com/2007/12/reducing-agi-complexity-copy-only-high.html
25. Fodor, J.: The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology, MIT Press, Cambridge, MA (2000)
26. Strong AI, Wikipedia, http://en.wikipedia.org/wiki/Strong_AI#Whole_brain_emulation
27. AI is NOT part of transhumanism, Human Enhancement and Biopolitics, http://hplusbiopolitics.wordpress.com/2008/06/13/ai-is-not-part-of-transhumanism/