RC24378 (W0710-075) October 15, 2007
Computer Science

IBM Research Report

Four Paths to AI

Jonathan Connell
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Kenneth Livingston
Psychology Department
Vassar College
Poughkeepsie, NY

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Four Paths to AI

Jonathan CONNELL[a] and Kenneth LIVINGSTON[b]
[a] *IBM T.J. Watson Research Center, Yorktown Heights NY, jconnell@us.ibm.com*
[b] *Psychology Department, Vassar College, Poughkeepsie NY, livingst@vassar.edu*

Abstract. There are a wide variety of approaches to Artificial Intelligence. Yet interestingly we find that these can all be grouped into four broad categories: Silver Bullets, Core Values, Emergence, and Emulation. We will explain the methodological underpinnings of these categories and give examples of the type of work being pursued in each. Understanding this spectrum of approaches can help defuse arguments between practitioners as well as elucidate common themes.

Keywords. Emergence, Animal Models, Consciousness, Language Understanding

1. Introduction – How can we achieve AI?

Artificial Intelligence has been pursued for over 40 years and has given rise to hundreds of different approaches. Early progress seemed rapid but halfway to Turing's goal of human-level AI the enterprise seemed to stall. In recent years a new generation of researchers has proposed a variety of ways to re-animate the search for general purpose AI. These proposals are diverse, and it is difficult to place bets on which approach might eventually prove successful, in large measure because the varied landscape of approaches is difficult to comprehend in a glance. We suggest, consistent with Lakatos's view of how scientific value is actually judged [1], that what is really needed is an effective way to catalog the different approaches. This then gives us a way to comprehend and evaluate the relative progress made by pursuing different approaches to building AI. This classification scheme can also help sort out whether an objection to some piece of work is directed at the technology itself, or rather at its methodological class. For instance, imagine a pacifist confronting a hawk. Instead baldly asserting that, "A gun will not solve your problem," a more constructive response would be, "Well, if you are going to fight, a gun is a reasonable weapon."

2. Silver Bullets – Just add missing mechanism X!

The approaches described here all have in common the suggestion that most of the necessary technology is already in place, we just need to resolve some particular nugget and then whole system will finally exhibit true intelligence. With this approach there is still some worry as to whether we have picked the right hole to fill. After all, a silver bullet will kill a werewolf, but not a vampire.

Fancy Logic – The idea is that first-order logic seems inadequate to the task of building AI, but that it can be achieved by moving to some more complex formal system of symbol manipulation. Techniques include various extensions of logic (e.g. second-order, non-monotonic, epistemic, deontic, modal, etc.) as well as other mechanisms like circumscription, abduction [2], and inductive logic programming.

Inexact Reasoning – The premise here is that formal symbol manipulation, like first-order logic, is too brittle for the real world. Things are not black-and-white but rather shades of gray, and AI systems need to be able to reason in this manner. Some interesting progress has been made using Fuzzy Logic for mobile robots [3].

Deep Language – An AI cannot be expected to be fully competent straight "out of the box", instead it needs to learn from sympathetic humans and/or from reading written material. To do this it must have a deep understanding of human language. To do this often involves a tight intermingling of syntactic and semantic [4].

Embodiment – Proponents of the embodiment solution to finding AI argue that you cannot achieve human-like intelligence unless the system has a body and can interact with the real physical world. Being embodied makes you care about objects, space, uncertainty, and actions to get tasks accomplished. In an important sense the body is just a special purpose computational engine, one that has evolved to solve very specific problems that are computationally expensive or even intractable any other way [5].

Quantum Physics – This line of argument suggests that consciousness is essential for true general intelligence, and that consciousness itself is based in quantum-level events. To achieve AI, therefore, will require finding ways to make quantum computing a reality. Although versions of the theory have been worked out in some detail as they might apply to the human case [6], the hypothesis has not been subjected to direct empirical test.

3. Core Values – Just make sure it can do X!

Much of this argument has to do with overall control structure, not specific types of computation. In fact, this approach argues that others are wrongheaded to concentrate on such details. If we just implement the correct central organizing principle everything else will fall into place. Yet such a strong core conceptualization brings its own vulnerabilities. Bad choices about the core principles can be disastrous because so much else builds from this core.

Situatedness – The reason none of our systems have achieved AI is they are not part of the real world – they do not have a gut feel for how interactions occur nor do they have a real stake in the outcome. This topic is concerned with the motivational structure of systems [7] as well as understanding physical and temporal processes.

Emotionality – Here the reasoning goes that emotion is not just a vestigial animal left-over or a mere a by-product of cognition but is instead an essential ingredient [8]. Emotion is crucial to regulating attention, mediating memory formation and retrieval, and arbitrating between actions in uncertain circumstances.

Self-Awareness – As a part of consciousness it is important to be able to recursively access information about one's own states. This gives a sense of a unitary self who deliberately chooses actions and experiences the benefits and costs of their consequences. It also forms the basis for predicting, imitating, and empathizing with other agents [9].

Hierarchy & Recursion –The ability to abstract from particulars to categorical representations is much more difficult than simple generalization. In fact, the ability to abstract recursively appears to be extremely rare and may even be limited to the human case [10]. The argument is that the most basic feature of general intelligence is a computational mechanism that takes any input, including its own outputs, and finds the pattern of differences and similarities that allow grouping into still more abstract categories [11]. This mechanism gives the data compression needed to produce meaningful but tractable understanding of very complex environments.

4. Emergence – Just add more X!

On this view, we actually have a pretty good grasp of the essentials but we haven't figured out how to implement them at the right scale. If we just add enough knowledge, speed, experience, etc. the system will "magically" come to life. This is a particularly popular mindset currently with the advent of fast processors, large memories, and so much machine-readable content online. Sometimes this strategy works well, as in the Deep Blue chess machine. It had a clever position evaluator, but the bulk of its strength came from a deep search of the game tree. Other times there is too much of an element of unreasoned faith involved. Galvani made a frog's leg twitch using a battery, so imagine (as Mary Shelley did) what a bolt of lightning would do!

Axiomatization – Classical first order logic underpins all human thought. It is a mere matter of identifying and formally codifying all the specific forms of reasoning and then writing the correct axioms for time, space, gravity, emotion, economics, social obligation, self-awareness, etc. There are a lot of these subfields to be encoded and this is the grist necessary for the mill of intelligence. This camp draws supporters from traditional logic backgrounds [12] as well as those working on Qualitative Physics.

Commonsense – This point of view says that we simply need to have the system understand the million or so facts that cover everyday existence. Most reasoning can then be based either directly on this pre-existing corpus, or on minimal extensions through analogy [13].

Learning – It is too hard (or even impossible) to program a creature to react appropriately in all situations. A more robust and flexible approach is to provide guidance about what are good situations versus bad ones and let it learn how to respond itself. All it needs is many time steps of experience in successively less sheltered environments. Reinforcement learning has been particularly successful here [14].

Evolution – This approach posits that the key to AI is self-improving systems. Even if the incremental steps are very small, as long as there is no theoretical bound then the system should be able to bootstrap its way to human-level performance (and beyond!). We just need lots of individuals and generations. Some interesting work has shown that physical structures [15] as well as control algorithms can be evolved.

Integration – A human is not just a brain in a box, it has eyes, ears, arms, legs, etc. How can an AI ever truly appreciate the meaning of a word like "red" without grounding it in some bodily sensation? We need to put everything we know how to do together in one place and let this creature experience the real physical world. The humanoid robot Cog [16] is one such ambitious attempt, but it is hard to have the best-of-breed technology in all categories simultaneously.

5. Emulation – Just faithfully copy X!

The emulation approach is pessimistic about whether we even have *any* of the proper mechanisms to create intelligence. Instead it advocates that existence proofs be copied. Technology often precedes science, so perhaps we can just re-implement some example in silicon and at least use it. Understanding and good theory can come later. Here simulating even (or especially) the faults of the underlying system is considered a virtue. Yet, since there is no underlying theory, it is hard to tell whether the details being copied are really relevant. For example, artificial feathers and flapping turn out not to be needed to create airplanes.

Neural simulation – All our computer metaphors for the brain may be entirely wrong. We need to simulate, as accurately as possible, the actual neural hardware and see how it responds to various stimuli. Without such detailed modeling (e.g. [17]) we may completely miss key aspects of how humans function.

Neural networks – The human mind presumably is a program that runs on hardware comprised of the human brain. However brains are organized very differently from standard digital computes, so perhaps starting with more a biologically-faithful substrate will make the AI problem easier. Particularly notable are subsymbolic approach to reasoning and language [18].

Animal models – Arguably humans evolved from "lower" animals and genetically the difference is quite small. This suggests that many of the mechanisms and behaviors present in animals underlie human intelligence and that the robust substrate provided by this heritage may be essential for cognition as we understand it. For instance, work on Skinner-bots [19] has shown how a robot can learn to fetch and recycle color objects in much the way a dog would be trained to do the same task.

Human development – How can we expect AI's to spring forth fully competent in several hours when infant development occurs over the course of many years? A lot is known about the various cognitive stages children progress through and it has been suggested that potential AI's follow this same developmental program [20].

Sociality – To be part of a larger cultural entity an AI needs to associate and communicate with other humans and robots [21]. To do this it needs to understand how to effectively participate in social interactions such as advice taking, negotiation, and collaboration. One of the most eye-catching projects here is the robot Kismet [22].

6. Discussion

Grouping things into categories, as in the periodic table of the elements, should serve to predict similar structure among entries in the same region of the table, as well as suggesting that one should observe related responses to various sorts of conditions. For instance, forgetting for the moment which secret ingredient is being promoted, is there any commonality about the "standard recipe" to which this ingredient will be added? Is it a symbolic substrate or a more diffuse set-based representation? When looking for a central organizing principle and asking the big questions, is the probe modality primarily verbal? Can relevant responses only be elicited in social situations? Similarly, for the emergent camp is there any way to predict how much of a resource will be needed to accomplish one task based on experience with another? Can we tell whether performance will asymptote (perhaps at an unacceptably low level) based on observed incremental improvement with added resource? And for emulation, how do we know whether a model is "faithful enough"? And are there any principles, even vague ones, pervading multiple types of emulation? Perhaps auto-encoders, entropy reduction, or reinforcement are recurring themes.

Taking this ten thousand foot view of the landscape it is even possible that insights gained along one path might have useful implications for another (e.g. the primacy of language, the necessity for task feedback). At the very least there is value in having a big picture view of where progress is being made and where it is stalled [1]. This lets us judge where resources of time and funding ought to be directed, and may be the closest thing available to an optimal search strategy for finding the right path or paths to AI.

References

[1] I. Lakatos, "Falsificationism and the Methodology of Scientific Research Programmes," in I. Lakatos and A. Musgrave (eds.), Criticism and the Growth of Knowledge, Cambridge University Press, 1965.

[2] J. Hobbs, M. Stickel, P. Martin, and D. Edwards, "Interpretation as Abduction", *Proc. Annual Mtg. of the Assoc. for Computational Linguistics*, pp. 95–103, 1990.

[3] A. Saffiotti, "Fuzzy Logic in Robot Navigation: A Case Study", Université Libre de Bruxelles, IRIDIA Tech Report 95-25, 1997.

[4] Michael Dyer, In-Depth Understanding, MIT Press, 1983.

[5] N. Kushmerick, "Software Agents and Their Bodies", *Minds and Machines*, vol. 7, pp. 227–247, 1997.

[6] N. Woolf and S. Hammeroff, "A Quantum Approach to Visual Consciousness*", TRENDS in Cognitive Sciences*, 5(11), pp. 472–478, 2001.

[7] S. Wilson, "Explore/Exploit Strategies in Autonomy", *From Animals to Animats* (Proc. SAB-96), 1996.

[8] Marvin Minsky, The Emotion Machine, Simon and Schuster, 2006.

[9] B. Scassellati, "Theory of Mind for a Humanoid Robot", *Autonomous Robots*, vol. 12, pp. 13–24, 2002.

[10] D. Premack, "Is Language the Key to Human Intelligence?", *Science*, 303(5656), pp. 318–320, 2004.

[11] Jeff Hawkins, On Intelligence, Owl Books, 2005.

[12] J. McCarthy, "The Well-Designed Child", http://www-formal.stanford.edu/jmc/child1.html, 1999.

[13] H. Liu and P. Singh, "ConceptNet – A Practical Commonsense Reasoning Toolkit", *BT Technology Journal*, 22(4), pp. 211–226, 2004.

[14] L. Lin, "Self-Improving Reactive Agents: Case Studies of Reinforcement Learning Frameworks", *From Animals to Animats* (Proc. SAB-90 Conf.), pp. 297–305, 1990.

[15] L. Lichtensteiger and P. Eggenberger, "Evolving the Morphology of a Compound Eye on a Robot", *Proc. of Eurobot*, pp. 127–134, 1999.

[16] R. Brooks, C. Breazeal, R. Irie, C. Kemp, M, Marjanović, B. Scassellati, and M. Williamson, "Alternative Essenses of Intelligence", Proc. of AAAI Conf., pp. 961–968, 1998.

[17] H. Markram, "The Blue Brain Project", *Nature Reviews Neuroscience*, vol. 7, pp. 153–160, 2006.

[18] R. Miikkulainen, "Natural Language Processing with Subsymbolic Neural Networks", in Neural Network Perspectives on Cognition and Adaptive Robotics, A. Browne (ed.), Taylor & Francis, 1997.

[19] L. Saksida, S. Raymond, and D. Touretsky, "Shaping Robot Behavior Using Principles from Instrumental Conditioning", *Robotics and Autonomous Systems*, 22 (3/4):231, 1998.

[20] A. Arsenio, "Children, Humanoid Robots and Caregivers", *Workshop on Epigenetic Robotics*, 2004.

[21] E. Herrmann, J. Call, M. Hernández-Lloreda, B. Hare, and M. Tomasello, "Humans Have Evolved Specialized Skills of Social Cognition", *Science*, vol. 317, pp. 1360–1366, 2007.

[22] Cynthia Breazeal, Designing Sociable Robots, MIT press, 2002.