

# Review of *Thinking about Android Epistemology*

Robert Morris

■ This article is a review of *Thinking about Android Epistemology* by Kenneth Ford, Patrick Hayes, and Clark Glymour. Cambridge, MA: AAAI Press/The MIT Press.

In the recent past, scientists have attempted to mimic conditions of the early universe at the atom smasher called RHIC (relativistic heavy ion collider) at the Brookhaven National Laboratory (Riorden and Zajc 2006). The result of colliding beams of gold traveling near the speed of light (“minibangs”) allows physicists to observe the liberation of quarks and gluons from protons and neutrons, revealing conditions that existed at the earliest moments of creation of the universe, thus validating current theories of how the original mix of quarks and gluons phase-transitioned into the mundane soup of protons and neutrons that forms the building blocks of everything. Theoretical and experimental breakthroughs since the 1970s, as well as technological advances in the art of colliding and detecting particles, have made it possible to observe a new “energy frontier,” with a wealth of results that will allow a refinement of our theories.

The question of the validity of the results obtained is a completely empirical matter. No one would seriously entertain the claim that the results obtained from RHIC are invalid because the results were obtained in an artificially induced laboratory setting rather than as the result of direct observation of nature. A (valid) simulation of the big bang is a (mini)-big bang.

By analogy, it is hard to observe directly in nature the mental states that lead to intelligent behavior because of the complexity of the brain and the lack of technology for examining these processes. Still, using symbolic representations of the building blocks of thinking, researchers in AI labs can execute programs that perform the sort of “symbol collision” that produces high-level thinking. Software systems that plan, control a complex device, or understand language simultaneously provide insight into the way such behavior is manifested in nature (namely, in human intelligence) and enable the development of automated software technologies for assisting humans in these tasks. As with cosmology, the results of artificially created intelligence validate theories of intelligence of the natural kind.

Android epistemology, as defined by the book under review, seeks to answer fundamental questions about the nature of such artificially intelligent machines. The theme of the book, “thinking about” machine intelligence, unfolds in a set of often entertaining essays by philosophers, cognitive scientists, and computer scientists. The topics related to this theme explored here are quite diverse and are typically presented in an informal, discursive style. Curious computer scientists specializing in AI will find the book useful for the purpose of surveying part of the philosophical underpinnings of AI. Especially insightful was the historical introduction by the book’s editors, which included a trace of the intellectual heritage of AI pioneers to philosopher mentors. Conversely, noncomputer scientists should have their minds expanded by essays such as those by Herbert Simon and Paul Churchland,



which clearly and concisely spell out and rigorously defend designs of architectures of machine intelligence.

The book is divided into four parts, each containing essays that roughly seek to answer the following four questions: (1) can machines be intelligent? (part I); (2) are human intelligence and machine intelligence based on the same underlying design principles? (part II); (3) what limitations, if any, to designing intelligent systems are provided by the frame problem? (part III); (4) what is the range of human traits that machines can exhibit? (part IV).

This book is a revision of the book *Android Epistemology*, published in 1995, containing a

mixture of new essays and ones from the earlier volume. The field of AI has significantly matured during the last 10 years, due in large measure to the concerted efforts to embed and integrate AI-based systems with non-AI software systems and hardware for use in robotic and other complex systems. This maturation has caused an expansion in focus with respect to aspects of rational agency being automated.

For a long time, the primary focus of AI was the goal-directedness of rational agents. That's why for quite a while the paradigm agent-based systems looked like planning systems (think of Shakey the robot). It is hard to imagine systems exhibiting robust goal-directed behavior that are not at least partially planners. But embedding AI into complex systems has led to the realization that rational agents are more than just goal-directed; they are also reactive, adaptive, mindful of the utility of their actions, and capable of learning models about the world (Russell and Norvig 2003). They use these capabilities not just to survive, but also to enhance their ability to accomplish goals. Only recently has the focus of AI shifted to these aspects of intelligence, led partially by Rodney Brooks's realization that nonplanning-based architectures for intelligence can be devised (Brooks 1991).

One minor problem of this book is that there are themes and remarks in it that seem to reflect the state of AI before the shift away from the focus on goal-directedness of agents. For example (p. 45): "We could build a device to recognize a voice, because sound patterns can be resolved by Fourier analysis and expressed mathematically. But faces? I wonder." I wonder whether the author possesses the same degree of pessimism today, given the advances made in face recognition (Mitchell 1997, chapter 4).

There are similarly other instances in the book where the skepticism is more about a state of the field of AI that may have changed. Similarly, there are other threads in the book that seem to reflect a too rigid set of assumptions about agent architectures—assumptions that Brooks and others have since challenged.

For example, Boden's discussion of creativity is built around a design for a constraint-based system that uses a heuristic to search a space of sequences (such as of sounds) for a "novel combination." Although perhaps a reasonable approach, the discussion here suffers from being rooted to a set of principles for designing intelligent agents that has been since expanded.

On the whole, however, this book contains much that is informative and visionary. Among the essays in part one, Clark Glymour's

entertaining “Silicon Reflections” shows by way of a clever fable that the claim that networks comprising “artificial sensory and motor nerves” cannot have mental states, that is, can’t think, feel, or understand, is hard to defend. The key is imagining an advance in medical technology whereby hybrid brains, part electromechanical, part brain matter, are possible. The underlying argument is a sort of Sorites paradox: if the result of replacing one brain cell in a brain with mental states with a mechanical equivalent is also a brain with mental states, then repeating this “operation” one more time should have the same effect; hence repeating it until the brain has completely been mechanized will produce something that has mental states. To avoid this conclusion, the “Dretskeans” (read: Searleans, deniers of the mechanical mind) are forced to either extreme or ad hoc positions. This essay also incorporates the theme, repeated in the closing chapter of the book and in other publications by the editors, that machine intelligence as a technology offers humans a sort of cognitive prosthesis—a way of augmenting the native capabilities of the human mind.

Essential reading for anyone interested in the foundations of machine intelligence is Herb Simon’s contribution. The language here is remarkably clear, lucid, and bold. It cuts through the rhetoric and nonsense that accompanied much of the debate around the Chinese room argument, giving each premise in the argument against AI its proper amount of space (which is often less than a sentence). Many of the themes Simon discusses here still make up the fundamental challenges for engineers of machine intelligence. Four of them include the following:

First is the focus on the response-time requirements of models that we build for decision-makers, devising concise, tractable representations of a complex search state for problem solving. Simon notes that the set of representations forms an ordered class on which notions of equivalence can be defined. He stresses the need for scalability and the importance of laboratory prototypes. It is clear that Simon always envisioned sophisticated agents observing and changing the world.

A second theme Simon discusses is the importance of “nearly decomposable” systems, which implies a layered architecture for control and deliberation with different levels of abstraction. Simon is clearly aware of the challenges of complexity in intelligent systems, and his comments about decomposable systems are also relevant to issues related to verification.

Third is the fact that processing in intelligent systems, whether human or machine, is distributed and parallel. Consequently, architectural issues of structure and “style” (how components interact) are important (Coste-Manire and Simmons 2000).

Finally is Simon’s discussion of reasoning with “ill-structured phenomena,” part of what today is called reasoning under uncertainty. Simon recognizes that imposing structure on ill-structured phenomena often forces a non-propositional representational framework. This insight is clearly reflected in the field of AI today.

Simon also boldly asserts that some arguments against machine intelligence are based on a failure to draw the proper distinctions between what is essential for mind versus what is not. In the latter category he discusses things like intention, consciousness, motivation, and awareness. Simon’s article offers a complete and general set of principles that form the underpinnings for an architecture of machine intelligence.

Another nice essay in part two is Paul Churchland’s technically detailed and crisp response to the charge that the content of consciousness cannot be mapped to an activation pattern in the brain because the latter differ between individuals whereas the former do not.

The best essays in part three are contributions by Daniel Dennett and Henry Kyburg. The frame problem, as Dennett notes, is an “installation problem,” a problem of creating a concise, finite model of action that can be used by an android to autonomously plan actions. The connection to autonomy is required; tele-operated systems or systems like the MER rovers, which are commanded remotely on the ground, do not suffer from the frame problem. Dennett speculates that the solution may reside in a shift in representational paradigm to something that would be referred to today as state-based planning. On this paradigm, an agent can be viewed as continuously observing the state of the world (a vector of values) and executing a policy on that state, construed as a function from states to actions. A policy can be viewed as a very large lookup table, and no enumeration of consequences of actions is ever required. Of course, devising a policy incurs its own technical challenges; the primary problem is the exponential size of the state space (in the number of variables). Indeed, the main challenge to such state-based approaches is in managing this complexity, but at the same time the frame problem dissolves. Dennett combines a serious discussion with playful stabs at academ-

ic philosophers, who emerge as simultaneously intellectually lazy (coming up with meaningful explanations are “not their problem”) and expert at pointing out the obvious.

Henry Kyburg picks up on many of Dennett’s themes in his contribution. His notion of practical certainty anticipates recent developments in probabilistic robotics (Thrun, Burgard, and Fox 2005). Specifically, his description of how beliefs are updated from new observations seems to map directly into what filtering algorithms do. Again, shifting the representational paradigm from propositions to one based on utilities dissolves the frame problem into a belief distribution.

Part four contains in general the weakest entries in the collection. The essay by Susan Sterrett proposes a variation of the Turing test for intelligence in terms of the ability of intelligent agents to “override instincts or habits.” The problem with this “test” is that it is clearly not empirically verifiable. A native Martian watching a MER rover navigate around a large rock might conclude it is overriding its habit of traveling in a straight line. The designers of the AutoNav system on MER (Maimones) would no doubt respond that its actions are completely habitual; faced with similar obstacles, the AutoNav system would always respond in the way observed.

A more interesting, but somewhat similar, test for consciousness was proposed by Christof Koch (Koch 2004, p. 227), who proposes an architecture for consciousness in the sensory cortex, and from this theory derives a test for consciousness roughly in terms of the ability of conscious agents not to respond behaviorally like “zombies.” If the theory is sound, truly conscious agents would adapt to changes to a stimulus after a short delay for processing, whereas agents with no consciousness would be incapable of adapting to the change. Koch’s test is more meaningful by virtue of its being an empirically verifiable test of a theoretical hypothesis.

Sterrett’s essay in general suffers from an obviously superficial understanding of AI architectures. The reader gets the sense of being invited to be impressed by the fact that an academic philosopher with very little technical knowledge of AI is able to come to grips successfully and accept the idea of machine intelligence.

In general, as noted by the authors, a deep technical knowledge of computer science, mathematics, or logic is not required to enjoy this “gentle introduction” to android epistemology. Nonetheless, many of the essays exhibit a sophistication, both in presentation

and content, that stems from the possession by the author of a deep underlying technical knowledge, and a reader of this collection is aided significantly by the possession of similar knowledge.

This book is a stimulating, fun read. Given the wide range of expertise of the contributors, as well as the range of attitudes exhibited in the essays, from light-hearted, even profane, to serious, it is unlikely that a reader will find all of the contributions of equal interest. Still, for researchers and students of AI, the essays by Simon, Churchland, Dennett, and Kyburg alone make this volume an essential contribution to the understanding of the principles that drive our pursuits.

## References

- Brooks, R. A. 1991. Intelligence without Representations. *Artificial Intelligence* 47(1–3): 139–159.
- Coste-Manire, E., and Simmons, R. 2000. Architecture: the Backbone of Robotic Systems. In *Proceedings of the 2000 IEEE International Conference on Robotics & Automation*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Koch, C. 2004. *The Quest for Consciousness: A Neurobiological Approach*. Greenwood Village, CO: Roberts and Company.
- Mitchell, T. 1997. *Machine Learning*. New York: McGraw Hill.
- Riordan, M., and Zajc, W. A. 2006. The First Few Microseconds. *Scientific American* 288(5) (May): 34–41.
- Russell, S., and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.
- Thrun, S.; Burgard, W.; and Fox, D. 2005. *Probabilistic Robotics*. Cambridge, MA: The MIT Press.



**Robert A. Morris** is a computer science researcher in the planning and scheduling group in the Intelligent Systems division at NASA Ames Research Center. He is part of NASA’s Intelligent Systems program, managing a number of projects that build automated reasoning systems for mission operations and autonomy. Currently he is technical lead in a project to develop an observation scheduling system for constellations of Earth-orbiting sensors. He received a B.A. from the University of Minnesota, a Ph.D. in philosophy from Indiana University, and an M.S. in computer science from Wright State University. The focus of his research is the area of temporal reasoning for planning and scheduling.