A Review of *Android Epistemology* by Kenneth M. Ford, Clark Glymour, and Patrick J. Hayes (eds) MIT Press, 1995 (To appear in *Robotika*)

Iain D. Craig
Department of Computer Science
University of Warwick
Coventry CV4 7AL
UK EU

September 17, 1996

A slightly edited version of this review is to appear in *Robotika*.

Android Epistemology is a collection of papers from a workshop held some years ago. It is notionally about the problem of knowledge for androids.

I confess immediately that writing this review has caused me considerable trouble. As will be seen, I disagree with most of the claims made by the editors, and I also disagree with many of the positions espoused by the various contributors; I also disagree profoundly with most of the assumptions made by the majority of the contributors and by the editors. The problem has been to produce a review of reasonable length without being unthinkingly negative and without being improper. The result is a review whose structure is a little odd for I want to outline the book and then discuss the contributions made by the editors. In outlining the book, I do not, however, want to summarise each chapter—that would take too long and would be extremely tedious.

Let me begin the outline by mentioning four chapters. First, there are the chapters by Simon and Boden. These merely repeat old themes. Simon bangs the symbol systems hypothesis drum again. The Boden again repeats her old story about creativity. The chapter by Churchland is about a neural network for visual perception; as such, it is of interest, but it is only obliquely a contribution to the problem of android epistemology (because it describes a computational model that performs a perceptual task).

The last chapter is by Minsky. It is a vaguely amusing dialogue between two hypothetical androids which discuss the differences between their construction and that of human beings. The discussion is not entirely restricted to serial machines, but it is clearly rooted in the current conceptions of what a computer is and what it can do. In mildly interesting way, it makes these differences explicit and points out those things that each does best. However, the language is somewhat grating (e.g., the self-ascription "that merry prankster"). The editors say in the *Introduction*:

> "We don't want to have dinner with anyone who reads it without laughing. We urge you to buy this book: Minsky's essay is worth the money, all by itself." (p. xvii)

As the reader will gather, I would *not* like to have dinner the editors. Moreover, this relatively slight chapter is not worth £20, nor is the rest of the book.

The rest of the book is highly variable in quality and value. There are chapters (those of Khan and Gips) on ethics and on the concept of the person (Gomila). There are chapters, on imagination (in the sense of the use of images) and subjective experience (Bringsjord and Stein), but they are somewhat disappointing. For example, Stein describes an addition to a robot which allows it to imagine future actions in a somewhat restricted form; given the title (*Imagination and Situated Cognition*), there is too little of a theoretical nature and too little discussion of how situatedness is involved in the processes of imagination. (The title of Stein's paper might cause confusion to some; the title uses the word 'imagination' in the sense of manipulating and using images rather than anything connected with inspiration—this is 'imagination' used in its proper sense and in the sense used in psychology.)

The theme of embodiment (for AI and computer people: embodiment is the idea that any 'intelligent' object must be possessed of a physical body in order to be 'intelligent'—Cognitive Sciences replace the term 'intelligent' by a complex of ideas) is also taken up by Chrisley in his chapter ('Taking embodiment seriously'). Chrisley considers the problem of thought in relation to sub- (or non-) conceptual content. Chrisley observes that there are significant problems with the specification of non-conceptual content (which one might take to be a specification in terms of sub-sentential constructs). The main choice is between two arguments, but there is ambivalence as to which account is the better, but Chrisley points to a review of the role of perception in the determination of content. This is an interesting paper, but, in my opinion, does not go far enough in asking what the relationship is between what happens inside the skin and what happens without.

I want to focus the rest of this review on the statements of the editors (I will refer to the editors of the book, who are also the authors of the *Introduction* and first chapter as 'the editors' in order to insist upon their identification) because they outline a position with which I disagree to a considerable extent, insisting as it does upon a particular perspective that depends upon a particular and extremely restricted view of the mental. The emphases and scope of the enterprise as seen by the editors exemplifies a particular bias. Indeed, at a recent summerschool, it was clear that a position close to that held by the editors is held by others, typically those whose work is closer to the 'Artificial Intelligence' (i.e. computer science) end of Cognitive Science, but there remains a risk that its influence will spread into other areas. Part of this appeal, it can be argued, is due to:

- A conception of meanings (and, hence, mental content) as *absolute* in some sense.
- A conception that only those processes which are purely 'mental' need be considered.
- A conception that the paradigm for whatever constitutes 'representation' is formal logic (whose meanings and contents are absolute in the sense of *a priori* given by the researcher/logician).

Not all of these beliefs can be ascribed to the contributors, but there appears to be a consensus, particularly with respect to the first to points (the issue of embodiment is also reduced to a mental/physical dichotomy in Chrisley's chapter).

This perspective is generally exhibited by the authors of the chapters of this book (although there is one clear counter-example—see below), as well as in the editors' choice of paper. The position is made amply clear as one reads the *Introduction* and the editors' chapter (the first, entitled 'The Prehistory of Android Epistemology'). The editors' view of the subject matter clearly determines their choice of material, and for that reason must be examined in some detail.

First, it must be noted that there are some rather off-putting and unfortunate remarks in the *Introduction* (the one about Minsky's chapter was mentioned above); a second, more vicious one, will be quoted below. The the third of these annoying comments is the following:

> "Were they reborn into a modern university, we claim, Plato, Aristotle and Leibniz would most suitably take up appointments in the department of computer science."

This unjustifiable remark seems far-fetched and overly parochial; furthermore, it falsely depicts the pre-occupations of the philosophers under consideration ("under selection" might be a better term), and it certainly misrepresents computer science as an exciting, dynamic, intellectually rich field of enquiry.

The chronological span of this chapter is of some concern, as are the emphases of the authors. This span starts with the Greeks and extends to Frege, Russell, and to Carnap's (1928) *Aufbau*. For the authors, philosophy stops at about 1930, the end of what is often referred to as the *Wienerkreis*. For the purposes of this book, philosophy ends with Logical Positivism, and ignores the insights of the later Wittgenstein, and of course ignores the work of Feyerabend, Heidegger and Husserl. What is presented is a distorted view of the history of philosophy which is manipulated to suit the needs of the authors in their belief that formal systems (such as formal logic) and the transduction and transformations of symbols of a kind familiar from formal logic lies at the core of an artificial intelligence (and presumably epistemology, philosophy of mind and the cognitive sciences).

Let us consider some of the examples from the history of philosophy which they cite and see why they are chosen.

The case of Ramón Lull (p. 6) is cited approvingly as an example of a philosopher who developed an abstract representation. Lull's 'represenation', if it can really be called such, was invented in order to *communicate* with non-christians; this fact is not considered by the editors, even though it can be convincingly argued that external representations such as this (and most of mathematics) are developed just so that concepts can be communicated. The editors would, no doubt, argue that abstract, symbolic representations are for *internal* use only (see p. 3, where they emphasise "a machine ... that could manipulate and be influenced by symbols"). Lull's contribution was more in the area of combinatorics, and was heavily influenced by the theological tradition to which he belonged.

Next, they consider Pascal, Leibniz and Bacon, collecting these three, very different, thinkers into one section (which, it is far too short—anything other than a book-length treatment of these themes would really be too short, it must be admitted). The section, like the one on the ancient Greeks, is highly selective. For example, there is no mention of Leibniz' work on monads or on causality, just as Aristotle is presented as being purely interested in empirical science and in logic. Here again, there is an attempt to fit history *backwards*; earlier, it is claimed that the Greeks invented *case-based reasoning*, and here the claim is that Bacon invented concept learning.

The next philosopher to be summarised is Descartes. Here, we hear of the Cartesian conception of mind which, they write "it is almost too banal to note, was of an entirely different substance than matter". It must be clear to anyone who is aware of the philosophy of mind that, this concept was so banal that it remains a significant problem in the philosophy of mind and has occupied the attentions of a number of great philosophers who are not mentioned by the editors (Husserl and Heidegger are the two primary examples from the twentieth century).

The section continues with two claims about Descarte's influence on 'modern' thought (I feel compelled to put the word 'modern' between quotation marks, for the view that anything in philosophy after 1928 is irrelevant does not, to my mind, present a 'modern' view of the subject). First, they claim that Descartes saw intention, meaning and rules as inextricably intertwined; this relates to their comment that duality "is tantalizingly close to the modern conception of software". However, throughout the entire 'history', we hear nothing about the criteria by which we judge rules to have been successfully applied, the nature of intentionality (is an intention ontological or epistemological in nature?), how meanings, and what the editors refer to as "representations", remain faithful to the world exterior to the observer, how it comes about that an essentially linguistic system has come to dominate other, evolutionarily prior processes (i.e., the representations assumed under the logicist approach, being logic-based, are essentially linguistic in nature). In short, there is a tacit assumption that everything is like logic.

The second aspect is that they state that the claim that what can be imagined must be possible leads to what they call the "Cartesian fallacy", citing Searle as one person who makes this error. However, they are also denying a general logical principle that if the positing of the negation of a proposition together with that proposition does not lead to contradiction, the negation is possible. Certainly, imagination is a weaker form of this, but if it is allowed that 'imagination' is taken in its widest sense, an important principle has been thrown away. Putnam (in a paper written long after 1928) has already shown that the status of the identity claims of modern science are open to question. (The section—pp. 12–13—tails off with another of those annoying sentences: "But before he could generate a single output Searle-in-the-box would become bored and his Chinese friends would have long since found better things to do.")

Next, we come to Hobbes and Kant. Hobbes is credited with anticipating Newell and Simon's physical symbol hypothesis; a lengthy quotation is aduced in the service of this claim (p. 13). On the way to Kant, Hume and Mill are given a sentence in which their connections with associationist psychology are mentioned. Then we come to Kant. Kant's problem with the *a priori* is glossed, and his theory of the 'intuition' is summarised in a sentence or two. The emphasis is on how Kant solved the problem of the *a priori* in terms of synthetic statements and in schematic thought which lead to the formation of a unified experience of the *a posteriori* given world.

And then to Boole, Frege and Freud. Boole invented an algebra of logic; Frege developed what was to become modern formal logic, but *Freud?!!!* Freud is mentioned in connection with D. O. Hebb's connectionist view of the mind (as propounded in *The Organization of Behvior*, 1949). Frued is then cited as a father figure for connectionism (p.17):

> But the basics of connectionist models of computation and of mind were fully developed by the late 19th century in the private views, and some of the public views, of neuropsychologists. We know of no better statement from this time that Sigmund Freud's private writings, where a great many details, including what is now called "Hebbian learning" and the "Hebbian synapse" are described."

But connectionism is clearly a different kind of thing from logic. This inclusion does not ring true; it is as if the editors wanted to placate the enemy with a kind word. The rest of the chapter is about logic, and stubbornly so.

Finally, we have a little eulogy. Hempel was Carnap's student, and Hempel's student Massey had a doctoral student called Buchanan who build MYCIN. MYCIN is therefore justified as falling within the canon of good taste.

The last sentence of the chapter is interesting (p. 20):

> "Philosophy should not be one more set of blinkers."

What the editors have done is to produce a blinkered, revisionist rewriting of the history of philosophy whose aim is to justify a particular and flawed approach to the problem of mind. In each case, they have selected parts of the work of the philosophes in question and have ignored others. For example, they ignore Aristotle's and Kant's work on ethics; they ignore most of Leibniz' work, and they ignore most people on causality and reference. Brentano ought to have been considered because it to him that we owe our modern conceptions of intentionality.

Furthermore, whole movements are ignored. Not only are the modern continental philosophers not mentioned, but the so-called Pragmatists (Peirce, Dewey, Mead and James) are not discussed (James gets a mention, but he described connectionism, or so they say). Putnam is another figure who deserves mention, if only for his work on identity statements.

The Pragmatists, of course, did not subscribe to a correspondence theory of truth, and the tacit subject of the chapter is the theories and thoughts that are claimed to have led to this conception of truth and Tarski's formalisaton of it within a particular kind of symbolic system. The editors do not mention the problem of reference, nor do they consider Davidson's analysis of meaning into truth and reference within a meta-linguistic system; an analysis of considerable importance to a Tarskian conception of meaning. What Davidson tries to do is to bridge the gap between purely formal systems and natural languages, and it is clear that something must be done in this respect, for logical systems on their own are insufficient. Indeed, it can be, and has been, argued by Barwise, *inter alia*, that the entire Tarskian enterprise is mistaken and that logical semantics in its current form is perfectly adequate for the foundations of mathematics, but ill-suited to the job of providing a semantics for natural language (and, by extension, to mental processes).

It can be seen that the major part of the first chapter is an attempt to trace the history of *symbolic* systems—i.e., logic—and to attempt to claim that logic and mind are inextricably intertwined, exactly as they claim for rules, meanings and intentions. The position adopted by Putnam is anti-realist, while the editors, again tacitly, assume some form of realism. The theme of the chapter is not a history of the ideas that lead to symbolic systems, but a justification of a particular hypothesis (one which has not been substantiated) and which presupposes a particular (and extremely questionable) conception of absolute truth (and hence absolute meaning).

Their view is, however, somewhat incoherent because they make the somewhat overblown claim (p. 4) that:

> "AI [Artificial Intelligence] has an engineer's respect for things that work

> and a pragmatic willingness to try anything that will produce the results needed. In the spirit of medieval astronomy, it wants only to save the phenomena, not to hypothesize psychological truth."

This pair of sentences suggest, first of all, that what is referred to as AI is in need of a coherent methodology, a methodology that can be used, *inter alia*, for the comparison of results (such comparisons are notoriously absent in AI), as well as for the direction of research activities. The second implication is that the activity is concerned with the production of artefacts of a particular kind, not the analysis of phenomena. At the same time, the quotation strongly implies that understanding of the phenomena is to be avoided at all costs; this is incoherent. By this reading, the only possible answer to the question as to the subject matter of 'AI' is the reply "I do not know". Even when a result is obtained in 'AI', one can not know whether it makes an empirical claim about the mind; this is denied by the editors' own claim. Furthermore, the entomological approach proposed is of necessity *pre*-theoretic, so one must question the purpose of a theoretical enterprise such as the current book.

Moreover, they are committed to a view which makes logical systems, and formally logical systems, the centre of attention. By means of these devices (whether on paper or implemented as programs) we can come to understand the mind.

A remark in the *Introduction* orients us to the editors' view on philosophy (p. xiv):

> We have included Kalyan Basu's essay in this collection for its forceful and compact statement of objections to android epistemology drawn from recent and contemporary English and German philosophy. We dispute virtually every unpoetic claim Basu makes for the singularity of human beings, but space and editorial politeness make us refrain from giving arguments here. We believe anyone reasonably informed about computer science and cognitive psychology will be able to provide some of them. Although none are directed explicitly to Basu's essay, the elements of many replies are contained in other chapters in this collection.

There appears within this statement to be an implicit claim that only the logical positivist approach (one which, incidentally, failed miserably in philosophy) cannot possibly provide insights.

Wittgenstein (in the *Philosophical Investigations*) makes the remark that "If a lion could speak, we could not understand him". This is the so-called *forms of life* argument in which Wittgenstein argues that the concepts and language uses of other species must differ from ours because of the fact that their perspective on the world is necessarily different from ours. This position is not simply an advanced (or alternative) form of the Whorffian hypothesis; it is a conclusion about the concepts that underpin the foundations of language and the conventions for their use.

Wittgenstein's arguments also pertain to the social structures in which we are submerged. Such concepts will not be acceptable to the editors of this book; they would claim that there is no reason to introduce such bogus concepts.

I do not recommend this book in any way. In my opinion, there are only a couple of chapters that are worth reading (and Basu's is the best in my opinion). The book does neither enlightens nor informs; nor did it inspire.

I believe it important to be as forcefully negative in this review as I have been. This book presents an incomplete picture and represents an unfortunately persuasive view of the field. Most of the conceptual content, eventually resting upon assumptions that have long been rejected as unhelpful or as inadequate, is to be rejected; the obsession with logical formalisms, it can be, and has been, argued (by many—including myself) leads nowhere. I consider it important to be as forceful as I have been for the reason that the arguments stated in this book are persuasive and convincing to those ignorant of philosphy (as most AI and computer people are); there are alternatives to the position adopted in *Android Epistemology* which make better sense and which avoid the fallacies inherent in a strange derivative of logical positivism.