Artificial Intelligence as Philosophy and as Psychology

Philosophers of mind have been interested in computers since their arrival a generation ago, but for the most part they have been interested only in the most abstract questions of principle, and have kept actual machines at arm's length and actual programs in soft focus. Had they chosen to take a closer look at the details I do not think they would have found much of philosophic interest until fairly recently, but recent work in Artificial Intelligence, or AI, promises to have a much more variegated impact on philosophy, and so, quite appropriately, philosophers have begun responding with interest to the bold manifestos of the Artificial Intelligentsia. My goal in this chapter is to provide a sort of travel guide to philosophers pursuing this interest. It is well known that amateur travelers in strange lands often ludicrously mis-comprehend what they see, and enthusiastically report wonders and monstrosities that later investigations prove never to have existed, while overlooking genuine novelties of the greatest importance. Having myself fallen prey to a variety of misconceptions about AI, and wasted a good deal of time and energy pursuing chimaeras, I would like to alert other philosophers to some of these pitfalls of interpretation. Since I am still acutely conscious of my own amateur status as an observer of AI, I must acknowledge at the outset that my vision of what is going on in AI, what is important and why, is almost certainly still somewhat untrustworthy. There is much in AI that I have not read, and much that I have read but not understood. So traveler, beware; take along any other maps you can find, and listen critically to the natives.

The interest of philosophers of mind in Artificial Intelligence comes as no surprise to many tough-minded experimental psychologists, for
from their point of view the two fields look very much alike: there are the same broad generalizations and bold extrapolations, the same blithe indifference to the hard-won data of the experimentalist, the same appeal to the deliverances of casual introspection and conceptual analysis, the aphoristic reasonings about what is impossible in principle or what must be the case in psychology. The only apparent difference between the two fields, such a psychologist might say, is that the AI worker pulls his armchair up to a console. I will argue that this observation is largely justified, but should not in most regards be viewed as a criticism. There is much work for the armchair psychologist to do, and a computer console has proven a useful tool in this work.

Psychology turns out to be very difficult. The task of psychology is to explain human perception, learning, cognition, and so forth in terms that will ultimately unite psychological theory to physiology in one way or another, and there are two broad strategies one could adopt: a bottom-up strategy that starts with some basic and well-defined unit or theoretical atom for psychology, and builds these atoms into molecules and larger aggregates that can account for the complex phenomena we all observe, or a top-down strategy that begins with a more abstract decomposition of the highest levels of psychological organization, and hopes to analyze these into more and more detailed smaller systems or processes until finally one arrives at elements familiar to the biologists. It is a commonplace that both endeavors could and should proceed simultaneously, but there is now abundant evidence that the bottom-up strategy in psychology is unlikely to prove very fruitful. The two best developed attempts at bottom-up psychology are stimulus-response behaviorism and what we might call "neuron signal physiological psychology", and both are now widely regarded as stymied, the former because stimuli and responses prove not to be perspicuously chosen atoms, the latter because synapses and impulse trains are perfectly good atoms, there are just too many of them, and their interactions are too complex to study once one abandons the afferent and efferent peripheries and tries to make sense of the crucial center (see Chapters 4 and 5).

Bottom-up strategies have not proved notably fruitful in the early development of other sciences, in chemistry and biology for instance, and so psychologists are only following the lead of "mature" sciences if they turn to the top-down approach. Within that broad strategy there are a variety of starting points that can be ordered in an array. Faced with the practical impossibility of answering the empirical questions of psychology by brute inspection (how in fact does the nervous system accomplish X or Y or Z?), psychologists ask themselves an easier preliminary question:
How could any system (with features A, B, C, . . .) possibly accomplish XI*

This sort of question is easier because it is "less empirical"; it is an engineering question, a quest for a solution (any solution) rather than a discovery. Seeking an answer to such a question can sometimes lead to the discovery of general constraints on all solutions (including of course nature’s as yet unknown solution), and therein lies the value of this style of aprioristic theorizing. Once one decides to do psychology this way, one can choose a degree of empirical difficulty for one’s question by filling in the blanks in the question schema above. The more empirical constraints one puts on the description of the system, or on the description of the requisite behavior, the greater the claim to "psychological reality" one’s answer must make. For instance, one can ask how any neuronal network with such-and-such physical features could possibly accomplish human color discriminations, or we can ask how any finite system could possibly subserve the acquisition of a natural language, or one can ask how human memory could possibly be so organized so as to make it so relatively easy for us to answer questions like "Have you ever ridden an antelope?", and so relatively hard to answer "What did you have for breakfast last Tuesday?". Or, one can ask, with Kant, how anything at all could possibly experience or know anything at all. Pure epistemology thus viewed, for instance, is simply the limiting case of the psychologists' quest, and is prima facie no less valuable to psychology for being so neutral with regard to empirical details. Some such questions are of course better designed to yield good answers than others, but properly carried out, any such investigation can yield constraints that bind all more data-enriched investigations.

AI workers can pitch their investigations at any level of empirical difficulty they wish; at Carnegie Mellon University, for instance, much is made of paying careful attention to experimental data on human performance, and attempting to model human performance closely. Other workers in AI are less concerned with that degree of psychological reality and have engaged in a more abstract version of AI. There is much that is of value and interest to psychology at the empirical end

*George Smith and Barbara Klein have pointed out to me that this question can be viewed as several ways ambiguous, and hence a variety of quite different responses might be held to answer such a question. Much of what I say below about different tactics for answering a question of this form can be construed to be about tactics for answering different (but related) questions. Philosophers who intend a question of this sort rhetorically can occasionally be embarrassed to receive in reply a detailed answer of one, variety of another.
of the spectrum, but I want to claim that AI is better viewed as sharing
with traditional epistemology the status of being a most general, most
abstract asking of the top-down question: how is knowledge possible?*
It has seemed to some philosophers that AI cannot be plausibly so
construed because it takes on an additional burden: it restricts itself to
mechanistic solutions, and hence its domain is not the Kantian domain
of all possible modes of intelligence, but just all possible mechanistically
realizable modes of intelligence. This, it is claimed, would beg the
question against vitalists, dualists and other anti-mechanists. But as I
have argued elsewhere, the mechanism requirement of AI is not an
additional constraint of any moment, for if psychology is possible at
all, and if Church's thesis is true, the constraint of mechanism is no
more severe than the constraint against begging the question in psy­
chology, and who would wish to evade that? (See Chapter 5).4

So I am claiming that AI shares with philosophy (in particular, with
epistemology and philosophy of mind) the status of most abstract
investigation of the principles of psychology. But it shares with psy­
chology in distinction from philosophy a typical tactic in answering its
questions. In AI or cognitive psychology the typical attempt to answer
a general top-down question consists in designing a particular system
that does, or appears to do, the relevant job, and then considering
which of its features are necessary not just to one's particular system
but to any such system. Philosophers have generally shunned such
elaborate system-designing in favor of more doggedly general inquiry.
This is perhaps the major difference between AI and "pure" philosopi­
cal approaches to the same questions, and it is one of my purposes here
to exhibit some of the relative strengths and weaknesses of the two
approaches.

The system-design approach that is common to AI and other styles
of top-down psychology is beset by a variety of dangers of which these
four are perhaps the chief:

(1) designing a system with component subsystems whose stipulated
capacities are miraculous given the constraints one is accepting. (E.g.,
positing more information-processing in a component than the
relevant time and matter will allow, or, at a more abstract level of
engineering incoherence, positing a subsystem whose duties would
require it to be more "intelligent" or "knowledgeable" than the
supersystem of which it is to be a part.

(2) mistaking conditional necessities of one's particular solution for
"This question (and attempts to answer it) constitutes one main branch of
epistemology: the other main branch has dealt with the problem of skepticism, and
its constitutive question might be: "Is knowledge possible?"
completely general constraints (a trivial example would be proclaiming that brains use LISP; less trivial examples require careful elucidation).

(3) restricting oneself artificially to the design of a subsystem (e.g., a depth perceiver or sentence parser) and concocting a solution that is systematically incapable of being grafted onto the other subsystems of a whole cognitive creature.

(4) restricting the performance of one's system to an artificially small part of the "natural" domain of that system and providing no efficient or plausible way for the system to be enlarged.

These dangers are altogether familiar to AI, but are just as common, if harder to diagnose conclusively, in other approaches to psychology. Consider danger (1): both Freud's ego subsystem and J.J. Gibson's invariance-sensitive perceptual "tuning forks" have been charged with miraculous capacities. Danger (2): behaviorists have been charged with illicitly extrapolating from pigeon-necessities to people-necessities, and it is often claimed that what the frog's eye tells the frog's brain is not at all what the person's eye tells the person's brain. Danger (3): it is notoriously hard to see how Chomsky's early syntax-driven system could interact with semantical components to produce or comprehend purposeful speech. Danger (4): it is hard to see how some models of nonsense-syllable rote memorization could be enlarged to handle similar but more sophisticated memory tasks. It is one of the great strengths of AI that when one of its products succumbs to any of these dangers this can usually be quite conclusively demonstrated.

I now have triangulated AI with respect to both philosophy and psychology (as my title suggested I would): AI can be (and should often be taken to be) as abstract and "unempirical" as philosophy in the questions it attempts to answer, but at the same time, it should be as explicit and particularistic in its models as psychology at its best. Thus one might learn as much of value to psychology or epistemology from a particular but highly unrealistic AI model as one could learn from a detailed psychology of, say, Martians. A good psychology of Martians, however unlike us they might be, would certainly yield general principles of psychology or epistemology applicable to human beings. Now before turning to the all important question: "What, so conceived, has AI accomplished?", I want to consider briefly some misinterpretations of AI that my sketch of it so far does not protect us from.

Since we are viewing AI as a species of top-down cognitive psychology, it is tempting to suppose that the decomposition of function in a computer is intended by AI to be somehow isomorphic to the decomposition of function in a brain. One learns of vast programs made up of
literally billions of basic computer events and somehow so organized as to produce a simulacrum of human intelligence, and it is altogether natural to suppose that since the brain is known to be composed of billions of tiny functioning parts, and since there is a gap of ignorance between our understanding of intelligent human behavior and our understanding of those tiny parts, the ultimate, millenial goal of AI must be to provide a hierarchical breakdown of parts in the computer that will mirror or be isomorphic to some hard-to-discover hierarchical breakdown of brain-event parts. The familiar theme of "organs made of tissues made of cells made of molecules made of atoms" is to be matched, one might suppose, in electronic hardware terms. In the thrall of this picture one might be discouraged to learn that some functional parts of the nervous system do not seem to function in the digital way the atomic functioning parts in computers do. The standard response to this worry would be that one had looked too deep in the computer (this is sometimes called the "grain problem")- The computer is a digital device at bottom, but a digital device can simulate an "analogue" device to any degree of continuity you desire, and at a higher level of aggregation in the computer one may find the analogue elements that are mappable onto the non-digital brain parts. As many writers have observed, we cannot gauge the psychological reality of a model until we are given the commentary on the model that tells us which features of the model are intended to mirror real saliencies in nature, and which are either backstage expediers of the modeling or sheer inadvertent detail. (In the Eighteenth Century, scientists built beautiful brass clockwork models of the solar system called orreries. The gears of an orrery are not intended to represent anything real in the firmament, nor are the reflections of one brass globe in another.) When one looks at AI programs in detail one sees mountains of digital calculation or "number crunching"; if this looks terribly unbiological, this is to be excused since that part of the model's structure is to be viewed as backstage, not part of the model we are supposed to compare with nature.

This story is fine in principle, I think, and the day may come when this imagined isomorphism down to the neuronal level (but not backstage) is proclaimed, but in fact we see nothing remotely like this in current AI work. We see instead models that diverge from nature very early on the downward path. AI programs typically model relatively high-level cognitive features, such as concept learning or revision of goals or beliefs, and it does indeed often take millions of basic computer events to simulate just a handful of these intended model features—but no one really knows or cares what these basic events are, for no one supposes that any of the saliencies in the mass of activity that
subserves the intended features are to be taken seriously as psychology or physiology. The psychological reality of an AI program is thus usually intended to be only skin deep; only the highest levels of program description and commentary are to be counted as the model.

The thinness of the psychological reality of AI models is obscured for the uninitiated by a particularly misleading feature of most of the AI programs that get into the limelight. What especially attracts audiences and evokes wonder is the marvelous verisimilitude of some of the programs. Consider this excerpt of conversation (minus the programmer’s excellent commentary) between Terry Winograd's SHRDLU and its interlocutor. SHRDLU is a program that lives in an imaginary "block world" where it can manipulate colored blocks on a table in response to commands and answer questions about its life in this world. Computer responses are capitalized.

Had you touched any pyramid before you put the green one on the little cube?

YES, THE GREEN ONE.

When did you pick it up?

WHILE I WAS STACKING UP THE RED CUBE, A LARGE RED BLOCK AND A LARGE GREEN CUBE.

Why?

TO GET RID OF IT.

Why did you do that?

TO CLEAR OFF THE RED CUBE.

Why did you clear off that cube?

TO PUT IT ON A LARGE GREEN CUBE.

Why did you do that?

BECAUSE YOU ASKED ME TO.6

When SHRDLU replies to the last question with "Because you asked me to", one can hardly help being struck by the apparent humanity of the response—"SHRDLU is one of us!" it is tempting to exclaim. If one is naive, one will be awestruck; if one is not, one will still be titillated by the illusion, for that is largely what it is. SHRDLU's response, though perfectly appropriate to the occasion (and not by coincidence!) is "canned". Winograd has simply given SHRDLU this whole sentence to print at times like these. If a child gave SHRDLU's
response we would naturally expect its behavior to manifest a general capacity which might also reveal itself by producing the response: "Because you told me to," or, "Because that's what I was asked to do," or on another occasion: "Because I felt like it," or "Because your assistant told me to," but these are dimensions of subtlety beyond SHRDLU. Its behavior is remarkably versatile, but it does not reveal a rich knowledge of interpersonal relations, of the difference between requests and orders, of being cooperative with other people under appropriate circumstances. (It should be added that Winograd's paper makes it very explicit where and to what extent he is canning SHRDLU's responses, so anyone who feels cheated by SHRDLU has simply not read Winograd. Other natural language programs do not rely on canned responses, or rely on them to a minimal extent.)

The fact remains, however, that much of the antagonism to AI is due to resentment and distrust engendered by such legerdemain. Why do AI people use these tricks? For many reasons. First, they need to get some tell-tale response back from the program and it is as easy to can a mnemonically vivid and "natural" response as something more sober, technical and understated (perhaps: "REASON: PRIOR COMMAND TO DO THAT"). Second, in Winograd's case he was attempting to reveal the minimal conditions for correct analysis of certain linguistic forms (note all the "problems" of pronominal antecedents in the sentences displayed), so "natural" language output to reveal correct analysis of natural language input was entirely appropriate. Third, AI people put canned responses in their programs because it is fun. It is fun to amuse one's colleagues, who are not fooled of course, and it is especially fun to bamboozle the outsiders. As an outsider, one must learn to be properly unimpressed by AI verisimilitude, as one is by the chemist's dazzling forest of glass tubing, or the angry mouths full of teeth painted on World War II fighter planes. (Joseph Weizenbaum's famous ELIZA program, the computer "psychotherapist" who apparently listens so wisely and sympathetically to one's problems, is intended in part as an antidote to the enthusiasm generated by AI verisimilitude. It is almost all clever canning, and is not a psychologically realistic model of anything, but rather a demonstration of how easily one can be gulled into attributing too much to a program. It exploits syntactic landmarks in one's input with nothing approaching genuine understanding, but it makes a good show of comprehension nevertheless. One might say it was a plausible model of a Wernicke's aphasic, who can babble on with well-formed and even semantically appropriate responses to his interlocutor, sometimes sustaining the illusion of comprehension for quite a while.)
The AI community pays a price for this misleading if fascinating fun, not only by contributing to the image of AI people as tricksters and hackers, but by fueling more serious misconceptions of the point of AI research. For instance, Winograd's real contribution in SHRDLU is not that he has produced an English speaker and understander that is psychologically realistic at many different levels of analysis (though that is what the verisimilitude strongly suggests, and what a lot of the fanfare—for which Winograd is not responsible—has assumed), but that he has explored some of the deepest demands on any system that can take direction (in a natural language), plan, change the world and keep track of the changes wrought or contemplated, and in the course of this exploration he has clarified the problems and proposed ingenious and plausible partial solutions to them. The real contribution in Winograd's work stands quite unimpeached by the perfectly true but irrelevant charge that SHRDLU doesn't have a rich or human understanding of most of the words in its very restricted vocabulary, or is terribly slow.

In fact, paying so much attention to the performance of SHRDLU (and similar systems) reveals a failure to recognize that AI programs are not empirical experiments but thought-experiments prosthetically regulated by computers. Some AI people have recently become fond of describing their discipline as "experimental epistemology". This unfortunate term should make a philosopher's blood boil, but if AI called itself thought-experimental epistemology (or even better: Gedanken-experimental epistemology) philosophers ought to be reassured. The questions asked and answered by the thought-experiments of AI are about whether or not one can obtain certain sorts of information processing-recognition, inference, control of various sorts, for instance—from certain sorts of designs. Often the answer is no. The process of elimination looms large in AI. Relatively plausible schemes are explored far enough to make it clear that they are utterly incapable of delivering the requisite behavior, and learning this is important progress, even if it doesn't result in a mind-boggling robot.

The hardware realizations of AI are almost gratuitous. Like dropping the cannonballs off the Leaning Tower of Pisa, they are demonstrations that are superfluous to those who have understood the argument, however persuasive they are to the rest. Are computers then irrelevant to AI? "In principle" they are irrelevant (in the same sense of "in principle", diagrams on the blackboard are in principle unnecessary to teaching geometry), but in practice they are not. I earlier described them as "prosthetic regulators" of thought-experiments. What I meant was this: it is notoriously difficult to keep wishful thinking out of one's thought-experiments; computer simulation forces one to recognize
all the costs of one's imagined design. As Pylyshyn observes, "What is
needed is ... a technical language with which to discipline one's
imagination." The discipline provided by computers is undeniable
(and especially palpable to the beginning programmer). It is both a
good thing—for the reasons just stated—and a bad thing. Perhaps you
have known a person so steeped in, say, playing bridge, that his entire
life becomes in his eyes a series of finesses, end plays and cross-ruffs.
Every morning he draws life's trumps and whenever he can see the end
of a project he views it as a lay-down. Computer languages seem to have
a similar effect on people who become fluent in them. Although I
won't try to prove it by citing examples, I think it is quite obvious that
the "technical language" Pylyshyn speaks of can cripple an imagination
in the process of disciplining it.10

It has been said so often that computers have huge effects on their
users' imaginations that one can easily lose sight of one of the most
obvious but still underrated ways in which computers achieve this
effect, and that is the sheer speed of computers. Before computers
came along the theoretician was strongly constrained to ignore the
possibility of truly massive and complex processes in psychology be­
cause it was hard to see how such processes could fail to appear at
worst mechanical and cumbersome, at best vegetatively slow, and of
course a hallmark of mentality is its swiftness. One might say that the
speed of thought defines the upper bound of subjective "fast", the
way the speed of light defines the upper bound of objective "fast".
Now suppose there had never been any computers but that somehow
(by magic, presumably) Kenneth Colby had managed to dream up these
flow charts as a proposed model of a part of human organization in
paranoia. (The flow charts are from his book, Artificial Paranoia,
Pergamon, 1975; figure 7.1 represents the main program; figures 7.2
and 7.3 are blow-ups of details of the main program.) It is obvious to
everyone, even Colby I think, that this is a vastly oversimplified model
of paranoia, but had there not been computers to show us how all this
processing and much much more can occur in a twinkling, we would
be inclined to dismiss the proposal immediately as altogether toe
clanking and inorganic, a Rube Goldberg machine. Most programs look
like that in slow motion (hand simulation) but speeded up they often
reveal a dexterity and grace that appears natural, and this grace is
entirely undetectable via a slow analysis of the program (cf. time lapse
photography of plants growing and buds opening). The grace in opera­
tion of AI programs may be mere illusion. Perhaps nature is graceful all
the way down, but for better or for worse, computer speed has
liberated the imagination of theoreticians by opening up the possibility
and plausibility of very complex interactive information processes playing a role in the production of cognitive events so swift as to be atomic to introspection.

At last I turn to the important question. Suppose that AI is viewed as I recommend, as a most abstract inquiry into the possibility of intelligence or knowledge. Has it solved any very general problems or discovered any very important constraints or principles? I think the answer is a qualified yes. In particular, I think AI has broken the back of an argument that has bedeviled philosophers and psychologists for over two hundred years. Here is a skeletal version of it: First, the only psychology that could possibly succeed in explaining the complexities of human activity must posit internal representations. This premise has been deemed obvious by just about everyone except the radical behaviorists (both in psychology and philosophy—both Watson and Skinner, and Ryle and Malcolm). Descartes doubted almost everything.
Figure 7-2
but this. For the British Empiricists, the internal representations were called ideas, sensations, impressions; more recently psychologists have talked of hypotheses, maps, schemas, images, propositions, engrams, neural signals, even holograms and whole innate theories. So the first premise is quite invulnerable, or at any rate it has an impressive mandate (see Chapter 6). But, second, nothing is intrinsically a representation of anything; something is a representation only for or to someone; any representation or system of representations thus requires at least one user or interpreter of the representation who is external to it. Any such interpreter must have a variety of psychological or intentional traits (see Chapter 1): it must be capable of a variety of comprehension, and must have beliefs and goals (so it can use the representation to inform itself and thus assist it in achieving its goals). Such an interpreter is then a sort of homunculus.

Therefore, psychology without homunculi is impossible. But psychology with homunculi is doomed to circularity or infinite regress, so psychology is impossible.

The argument given is a relatively abstract version of a familiar group of problems. For instance, it seems (to many) that we cannot account for perception unless we suppose it provides us with an internal image (or model or map) of the external world, and yet what good would that image do us unless we have an inner eye to perceive it, and how are we to explain its capacity for perception? It also seems (to many) that understanding a heard sentence must be somehow translating it into some internal message, but how will this message in turn be understood: by translating it into something else? The problem is an old one, and let's call it Hume's Problem, for while he did not state it explicitly, he appreciated its force and strove mightily to escape its clutches. Hume's internal representations were impressions and ideas, and he wisely shunned the notion of an inner self that would intelligently manipulate these items, but this left him with the necessity of getting the ideas and impressions to "think for themselves". The result was his theory of the self as a "bundle" of (nothing but) impressions and ideas. He attempted to set these impressions and ideas into dynamic interaction by positing various associationistic links, so that each succeeding idea in the stream of consciousness dragged its successor onto the stage according to one or another principle, all without benefit of intelligent supervision. It didn't work, of course. It couldn't conceivably work, and Hume's failure is plausibly viewed as the harbinger of doom for any remotely analogous enterprise. On the one hand, how could any theory of psychology make sense of representations that understand themselves, and on the other, how could any
theory of psychology avoid regress or circularity if it posits at least one representation-understander in addition to the representations?

Now no doubt some philosophers and psychologists who have appealed to internal representations over the years have believed in their hearts that somehow the force of this argument could be blunted, that Hume’s problem could be solved, but I am sure no one had the slightest idea how to do this until AI and the notion of data-structures came along. Data-structures may or may not be biologically or psychologically realistic representations, but they are, if not living, breathing examples, at least clanking, functioning examples of representations that can be said in the requisite sense to understand themselves.*

How this is accomplished can be metaphorically described (and any talk about internal representations is bound to have a large element of metaphor in it) by elaborating our description (see Chapter 5) of AI as a top-down theoretical inquiry. One starts, in AI, with a specification of a whole person or cognitive organism—what I call, more neutrally, an intentional system (see Chapter 1)—or some artificial segment of that person’s abilities (e.g., chess-playing, answering questions about baseball) and then breaks that largest intentional system into an organization of subsystems, each of which could itself be viewed as an intentional system (with its own specialized beliefs and desires) and hence as formally a homunculus. In fact, homunculus talk is ubiquitous in AI, and almost always illuminating. AI homunculi talk to each other, wrest control from each other, volunteer, sub-contract, supervise, and even kill. There seems no better way of describing what is going on.** Homunculi are bogeymen only if they duplicate entire the talents they are rung in to explain (a special case of danger (1)). If one can get a team or committee of relatively ignorant, narrow-minded, blind homunculi to produce the intelligent behavior of the whole, this is progress. A flow chart is typically the organizational chart of a committee of homunculi (investigators, librarians, accountants, executives); each box specifies a homunculus by prescribing a function without

*Joseph Weizenbaum has pointed out to me that Turing saw from the very beginning that computers could in principle break the threatened regress of Hume’s Problem, and George Smith has drawn my attention to similar early wisdom in Von Neumann. It has taken a generation of development for their profound insights to be confirmed, after a fashion, by detailed models. It is one thing—for from negligible—to proclaim a possibility in principle, and another to reveal how the possibility might be made actual in detail. Before the relatively recent inventions of AI, the belief that Hume’s Problem could be dissolved somehow by the conceptual advances of computer science provided encouragement but scant guidance to psychologists and philosophers.
saying how it is to be accomplished (one says, in effect: put a little man in there to do the job). If we then look closer at the individual boxes we see that the function of each is accomplished by subdividing it via another flow chart into still smaller, more stupid homunculi. Eventually this nesting of boxes within boxes lands you with homunculi so stupid (all they have to do is remember whether to say yes or no when asked) that they can be, as one says, "replaced by a machine". One discharges fancy homunculi from one's scheme by organizing armies of such idiots to do the work.

When homunculi at a level interact, they do so by sending messages, and each homunculus has representations that it uses to execute its functions. Thus typical AI discussions do draw a distinction between representation and representation-user: they take the first step of the threatened infinite regress, but as many writers in AI have observed, it has gradually emerged from the tinkerings of AI that there is a trade-off between sophistication in the representation and sophistication in the user. The more raw and uninterpreted the representation—e.g., the mosaic of retinal stimulation at an instant—the more sophisticated the interpreter or user of the representation. The more interpreted a representation—the more procedural information is embodied in it, for instance—the less fancy the interpreter need be. It is this fact that permits one to get away with lesser homunculi at high levels, by getting their earlier or lower brethren to do some of the work. One never quite gets completely self-understanding representations (unless one stands back and views all representation in the system from a global vantage point), but all homunculi are ultimately discharged. One gets the advantage of the trade-off only by sacrificing versatility and universality in one's subsystems and their representations, so one's homunculi cannot be too versatile nor can the messages they send and receive have the full flavor of normal human linguistic interaction. We have seen an example of how homuncular communications may fall short in SHRDLU's remark, "Because you asked me to." The context of production and the function of the utterance makes clear that this is a sophisticated communication and the product of a sophisticated representation, but it is not a full-fledged Gricean speech act. If it were, it would require too fancy a homunculus to use it.

There are two ways a philosopher might view AI data structures. One could grant that they are indeed self-understanding representations or one could cite the various disanalogies between them and prototypical or real representations (human statements, paintings, maps) and conclude that data-structures are not really internal representations at all. But if one takes the latter line, the modest successes of AI simply serve
to undercut our first premise: it is no longer obvious that psychology needs internal representations; internal pseudo-representations may do just as well.

It is certainly tempting to argue that since AI has provided us with the only known way of solving Hume's Problem, albeit for very restrictive systems, it must be on the right track, and its categories must be psychologically real, but one might well be falling into Danger (2) if one did. We can all be relieved and encouraged to learn that there is a way of solving Hume's Problem, but it has yet to be shown that AI's way is the only way it can be done.

AI has made a major contribution to philosophy and psychology by revealing a particular way in which simple cases of Hume's Problem can be solved. What else has it accomplished of interest to philosophers? I will close by just drawing attention to the two main areas where I think the AI approach is of particular relevance to philosophy.

For many years philosophers and psychologists have debated (with scant interdisciplinary communication) about the existence and nature of mental images. These discussions have been relatively fruitless, largely, I think, because neither side had any idea of how to come to grips with Hume's Problem. Recent work in AI, however, has recast the issues in a clearly more perspicuous and powerful framework, and anyone hoping to resolve this ancient issue will find help in the AI discussions.¹⁵

The second main area of philosophical interest, in my view, is the so-called "frame problem."¹⁶ The frame problem is an abstract epistemological problem that was in effect discovered by AI thought-experimentation. When a cognitive creature, an entity with many beliefs about the world, performs an act, the world changes and many of the creature's beliefs must be revised or updated. How? It cannot be that we perceive and notice all the changes (for one thing, many of the changes we know to occur do not occur in our perceptual fields), and hence it cannot be that we rely entirely on perceptual input to revise our beliefs. So we must have internal ways of up-dating our beliefs that will fill in the gaps and keep our internal model, the totality of our beliefs, roughly faithful to the world.

If one supposes, as philosophers traditionally have, that one's beliefs are a set of propositions, and reasoning is inference or deduction from members of the set, one is in for trouble, for it is quite clear (though still controversial) that systems relying only on such processes get swamped by combinatorial explosions in the updating effort. It seems that our entire conception of belief and reasoning must be radically revised if we are to explain the undeniable capacity of
human beings to keep their beliefs roughly consonant with the reality they live in.

I think one can find an appreciation of the frame problem in Kant (we might call the frame problem Kant's Problem) but unless one disciplines one's thought-experiments in the AI manner, philosophical proposals of solutions to the problem, including Kant's of course, can be viewed as at best suggestive, at worst mere wishful thinking.

I do not want to suggest that philosophers abandon traditional philosophical methods and retrain themselves as AI workers. There is plenty of work to do by thought-experimentation and argumentation, disciplined by the canons of philosophical method and informed by the philosophical tradition. Some of the most influential recent work in AI (e.g., Minsky's papers on "Frames") is loaded with recognizably philosophical speculations of a relatively unsophisticated nature. Philosopher's wheels, on the other hand, are perfect circles, require in principle no lubrication, and can go in at least two directions at once. Clearly a meeting of minds is in order.*

*I am indebted to Margaret Boden for valuable advice on an early draft of this paper. Her Artificial Intelligence and Natural Man (Harvester, 1977), provides an excellent introduction to the field of AI for philosophers.