Dennett, Daniel, C (September 1-3, 1994) . Consciousness in Human and Robot Minds. IIAS Symposium on Cognition, Computation and Consciousness.

Consciousness in Human and Robot Minds

[For IIAS Symposium on Cognition, Computation and Consciousness, Kyoto, September 1-3, 1994, forthcoming in Ito, et al., eds., *Cognition, Computation and Consciousness*, OUP. An earlier version of this paper was presented to the Royal Society, London, April 14, 1994.]

Daniel C. Dennett
Center for Cognitive Studies
Tufts University
Medford, MA, USA

1. Good and Bad Grounds for Skepticism

The best reason for believing that robots might some day become conscious is that we human beings are conscious, and we are a *sort* of robot ourselves. That is, we are extraordinarily complex self-controlling, self-sustaining physical mechanisms, designed over the eons by natural selection, and operating according to the same well-understood principles that govern all the other physical processes in living things: digestive and metabolic processes, self-repair and reproductive processes, for instance. It may be wildly over-ambitious to suppose that human artificers can repeat Nature's triumph, with variations in material, form, and design process, but this is not a deep objection. It is not as if a conscious machine contradicted any fundamental laws of nature, the way a perpetual motion machine does. Still, many skeptics believe--or in any event want to believe--that it will never be done. I wouldn't wager against them, but my reasons for skepticism are mundane, economic reasons, not theoretical reasons.

Conscious robots probably will always simply cost too much to make. Nobody will ever synthesize a gall bladder out of atoms of the requisite elements, but I think it is uncontroversial that a gall bladder is nevertheless "just" a stupendous assembly of such atoms. Might a conscious robot be "just" a stupendous assembly of more elementary artifacts--silicon chips, wires, tiny motors and cameras--or would any such assembly, of whatever size and sophistication, have to leave out some special ingredient that is requisite for consciousness?

Let us briefly survey a nested series of reasons someone might advance for the impossibility of a conscious robot:

> (1) Robots are purely material things, and consciousness requires immaterial mind-stuff. (Old-fashioned dualism)

It continues to amaze me how attractive this position still is to many people. I would have thought a historical perspective alone would make this view seem ludicrous: over the centuries, every *other* phenomenon of initially "supernatural" mysteriousness has succumbed to an uncontroversial explanation

within the commodious folds of physical science. Thales, the Pre-Socratic proto-scientist, thought the loadstone had a soul, but we now know better; magnetism is one of the best understood of physical phenomena, strange though its manifestations are. The "miracles" of life itself, and of reproduction, are now analyzed into the well-known intricacies of molecular biology. Why should consciousness be any exception? Why should the brain be the only complex physical object in the universe to have an interface with another realm of being? Besides, the notorious problems with the supposed transactions at that dualistic interface are as good as a *reductio ad absurdum* of the view. The phenomena of consciousness are an admittedly dazzling lot, but I suspect that dualism would never be seriously considered if there weren't such a strong undercurrent of desire to protect the mind from science, by supposing it composed of a stuff that is in principle uninvestigatable by the methods of the physical sciences.

But if you are willing to concede the hopelessness of dualism, and accept some version of materialism, you might still hold:

> (2) Robots are inorganic (by definition), and consciousness can exist only in an organic brain.

Why might this be? Instead of just hooting this view off the stage as an embarrassing throwback to old-fashioned vitalism, we might pause to note that there is a respectable, if not very interesting, way of defending this claim. Vitalism is deservedly dead; as biochemistry has shown in matchless detail, the powers of organic compounds are themselves all mechanistically reducible and hence mechanistically reproducible at one scale or another in alternative physical media; but it is conceivable--if unlikely--that the sheer speed and compactness of biochemically engineered processes in the brain are in fact unreproducible in other physical media (Dennett, 1987). So there might be straightforward reasons of engineering that showed that any robot that could not make use of organic tissues of one sort or another within its fabric would be too ungainly to execute some task critical for consciousness. If making a conscious robot were conceived of as a sort of sporting event--like the America's Cup--rather than a scientific endeavor, this could raise a curious conflict over the official rules. Team A wants to use artificially constructed organic polymer "muscles" to move its robot's limbs, because otherwise the motor noise wreaks havoc with the robot's artificial ears. Should this be allowed? Is a robot with "muscles" instead of motors a robot within the meaning of the act? If muscles are allowed, what about lining the robot's artificial retinas with genuine organic rods and cones instead of relying on relatively clumsy color-tv technology?

I take it that no serious scientific or philosophical thesis links its fate to the fate of the proposition that a *protein-free* conscious robot can be made, for example. The standard understanding that a robot shall be made of metal, silicon chips, glass, plastic, rubber and such, is an expression of the willingness of theorists to bet on a simplification of the issues: their conviction is that the crucial functions of intelligence can be achieved by one high-level simulation or another, so that it would be no undue hardship to restrict themselves to these materials, the readily available cost-effective ingredients in any case. But if somebody were to invent some sort of cheap artificial neural network fabric that could usefully be spliced into various tight corners in a robot's control system, the embarrassing fact that this fabric was made of organic molecules would not and should not dissuade serious roboticists from using it--and simply taking on the burden of explaining to the uninitiated why this did not constitute "cheating" in any important sense.

I have discovered that some people are attracted by a third reason for believing in the impossibility of conscious robots.

> (3) Robots are artifacts, and consciousness abhors an artifact; only something natural, born not manufactured, could exhibit genuine consciousness.

Once again, it is tempting to dismiss this claim with derision, and in some of its forms, derision is just what it deserves. Consider the general category of creed we might call *origin essentialism:* only wine made under

the direction of the proprietors of Chateau Plonque counts as genuine Chateau Plonque; only a canvas every blotch on which was caused by the hand of Cezanne counts as a genuine Cezanne; only someone "with Cherokee blood" can be a real Cherokee. There are perfectly respectable reasons, eminently defensible in a court of law, for maintaining such distinctions, so long as they are understood to be protections of rights growing out of historical processes. If they are interpreted, however, as indicators of "intrinsic properties" that set their holders apart from their otherwise indistinguishable counterparts, they are pernicious nonsense. Let us dub *origin chauvinism* the category of view that holds out for some mystic difference (a difference of value, typically) due *simply* to such a fact about origin. Perfect imitation Chateau Plonque is exactly as good a wine as the real thing, counterfeit though it is, and the same holds for the fake Cezanne, if it is really indistinguishable by experts. And of course no person is intrinsically better or worse in any regard just for having or not having Cherokee (or Jewish, or African) "blood."

And to take a threadbare philosophical example, an atom-for-atom duplicate of a human being, an artifactual counterfeit of you, let us say, might not *legally* be you, and hence might not be entitled to your belongings, or deserve your punishments, but the suggestion that such a being would not be a feeling, conscious, alive *person* as genuine as any born of woman is preposterous nonsense, all the more deserving of our ridicule because if taken seriously it might seem to lend credibility to the racist drivel with which it shares a bogus "intuition".

If consciousness abhors an artifact, it cannot be because being born gives a complex of cells a property (aside from that historic property itself) that it could not otherwise have "in principle". There might, however, be a question of practicality. We have just seen how, as a matter of exigent practicality, it could turn out after all that organic materials were needed to make a conscious robot. For similar reasons, it could turn out that any conscious robot had to be, if not born, at least the beneficiary of a longish period of infancy. Making a fully-equipped conscious adult robot might just be too much work. It might be vastly easier to make an initially unconscious or nonconscious "infant" robot and let it "grow up" into consciousness, more or less the way we all do. This hunch is not the disreputable claim that a certain sort of historic process puts a mystic stamp of approval on its product, but the more interesting and plausible claim that a certain sort of process is the only practical way of designing all the things that need designing in a conscious being.

Such a claim is entirely reasonable. Compare it to the claim one might make about the creation of Steven Spielberg's film, *Schindler's List*: it could not have been created entirely by computer animation, without the filming of real live actors. This impossibility claim must be false "in principle," since every frame of that film is nothing more than a matrix of gray-scale pixels of the sort that computer animation can manifestly create, at any level of detail or "realism" you are willing to pay for. There is nothing mystical, however, about the claim that it would be practically impossible to render the nuances of that film by such a bizarre exercise of technology. How much easier it is, practically, to put actors in the relevant circumstances, in a concrete simulation of the scenes one wishes to portray, and let them, via ensemble activity and re-activity, provide the information to the cameras that will then fill in all the pixels in each frame. This little exercise of the imagination helps to drive home just how much information there is in a "realistic" film, but even a great film, such as *Schindler's List*, for all its complexity, is a simple, non-interactive artifact many orders of magnitude less complex than a conscious being.

When robot-makers have claimed in the past that in principle they could construct "by hand" a conscious robot, this was a hubristic overstatement analogous to what Walt Disney might once have proclaimed: that his studio of animators could create a film so realistic that no one would be able to tell that it was a cartoon, not a "live action" film. What Disney couldn't do in fact, computer animators still cannot do, but perhaps only for the time being. Robot makers, even with the latest high-tech innovations, also fall far short of their hubristic goals, now and for the foreseeable future. The comparison serves to expose the likely source of the outrage so many skeptics feel when they encounter the manifestos of the Artificial Intelligencia. Anyone

who seriously claimed that *Schindler's List* could in fact have been made by computer animation could be seen to betray an obscenely impoverished sense of what is conveyed in that film. An important element of the film's power if the fact that it *is* a film made by assembling human actors to portray those events, and that it is not actually the newsreel footage that its black-and-white format reminds you of. When one juxtaposes in one's imagination a sense of what the actors must have gone through to make the film with a sense of what the people who actually lived the events went through, this reflection sets up reverberations in one's thinking that draw attention to the deeper meanings of the film. Similarly, when robot enthusiasts proclaim the likelihood that they can simply *construct* a conscious robot, there is an understandable suspicion that they are simply betraying an infantile grasp of the subtleties of conscious life. (I hope I have put enough feeling into that condemnation to satisfy the skeptics.)

But however justified that might be in some instances as an *ad hominem* suspicion, it is simply irrelevant to the important theoretical issues. Perhaps no cartoon could be a great film, but they are certainly real films--and some are indeed good films; if the best the roboticists can hope for is the creation of some crude, cheesy, second-rate, artificial consciousness, they still win. Still, it is not a foregone conclusion that even this modest goal is reachable. If you want to have a defensible reason for claiming that no conscious robot will ever be created, you might want to settle for this:

(4) Robots will always just be much too simple to be conscious.

After all, a normal human being is composed of trillions of parts (if we descend to the level of the macromolecules), and many of these rival in complexity and design cunning the fanciest artifacts that have ever been created. We consist of billions of cells, and a single human cell contains within itself complex "machinery" that is still well beyond the artifactual powers of engineers. We are composed of thousands of different kinds of cells, including thousands of different species of symbiont visitors, some of whom might be as important to our consciousness as others are to our ability to digest our food! If all that complexity were needed for consciousness to exist, then the task of making a single conscious robot would dwarf the entire scientific and engineering resources of the planet for millennia. And who would pay for it?

If no other reason can be found, this may do to ground your skepticism about conscious robots in your future, but one shortcoming of this last reason is that it is scientifically boring. If this is the only reason there won't be conscious robots, then consciousness isn't that special, after all. Another shortcoming with this reason is that it is dubious on its face. Everywhere else we have looked, we have found higher-level commonalities of function that permit us to substitute relatively simple bits for fiendishly complicated bits. Artificial heart valves work really very well, but they are orders of magnitude simpler than organic heart valves, heart valves born of woman or sow, you might say. Artificial ears and eyes that will do a serviceable (if crude) job of substituting for lost perceptual organs are visible on the horizon, and anyone who doubts they are possible in principle is simply out of touch. Nobody ever said a prosthetic eye had to see as keenly, or focus as fast, or be as sensitive to color gradations as a normal human (or other animal) eye in order to "count" as an eye. If an eye, why not an optic nerve (or acceptable substitute thereof), and so forth, all the way in?

Some (Searle, 1992, Mangan, 1993) have supposed, most improbably, that this proposed regress would somewhere run into a non-fungible medium of consciousness, a part of the brain that could not be substituted on pain of death or zombiehood. Once the implications of that view are spelled out (Dennett, 1993a, 1993b), one can see that it is a non-starter. There is no reason at all to believe that some one part of the brain is utterly irreplacible by prosthesis, provided we allow that some crudity, some loss of function, is to be expected in most substitutions of the simple for the complex. An artificial brain is, on the face of it, as "possible in principle" as an artificial heart, just much, much harder to make and hook up. Of course once we start letting crude forms of prosthetic consciousness--like crude forms of prosthetic vision or hearing--pass our litmus tests for consciousness (whichever tests we favor) the way is open for another boring debate,

over whether the phenomena in question are too crude to count.

2. The Cog Project: A Humanoid Robot

A much more interesting tack to explore, in my opinion, is simply to set out to make a robot that is theoretically interesting independent of the philosophical conundrum about whether it is conscious. Such a robot would have to perform a lot of the feats that we have typically associated with consciousness in the past, but we would not need to dwell on that issue from the outset. Maybe we could even learn something interesting about what the truly hard problems are without ever settling any of the issues about consciousness.

Such a project is now underway at MIT. Under the direction of Professors Rodney Brooks and Lynn Andrea Stein of the AI Lab, a group of bright, hard-working young graduate students are laboring as I speak to create Cog, the most humanoid robot yet attempted, and I am happy to be be a part of the Cog team. Cog is just about life-size--that is, about the size of a human adult. Cog has no legs, but lives bolted at the hips, you might say, to its stand. It has two human-length arms, however, with somewhat simple hands on the wrists. It can bend at the waist and swing its torso, and its head moves with three degrees of freedom just about the way yours does. It has two eyes, each equipped with both a foveal high-resolution vision area and a low-resolution wide-angle parafoveal vision area, and these eyes saccade at almost human speed. That is, the two eyes can complete approximately three fixations a second, while you and I can manage four or five. Your foveas are at the center of your retinas, surrounded by the grainier low-resolution parafoveal areas; for reasons of engineering simplicity, Cog's eyes have their foveas mounted above their wide-angle vision areas.

This is typical of the sort of compromise that the Cog team is willing to make. It amounts to a wager that a vision system with the foveas moved out of the middle can still work well enough not to be debilitating, and the problems encountered will not be irrelevant to the problems encountered in normal human vision. After all, nature gives us examples of other eyes with different foveal arrangements. Eagles have three different foveas in each eye, for instance, and rabbit eyes are another story all together. Cog's eyes won't give it visual information exactly like that provided to human vision by human eyes (in fact, of course, it will be vastly degraded), but the wager is that this will be plenty to give Cog the opportunity to perform impressive feats of hand-eye coordination, identification, and search. At the outset, Cog will not have color vision.

Since its eyes are video cameras mounted on delicate, fast-moving gimbals, it might be disastrous if Cog were inadvertently to punch itself in the eye, so part of the hard-wiring that must be provided in advance is an "innate" if rudimentary "pain" or "alarm" system to serve roughly the same protective functions as the reflex eye-blink and pain-avoidance systems hard-wired into human infants.

Cog will not be an adult at first, in spite of its adult size. It is being designed to pass through an extended period of artificial infancy, during which it will have to learn from experience, experience it will gain in the rough-and-tumble environment of the real world. Like a human infant, however, it will need a great deal of protection at the outset, in spite of the fact that it will be equipped with many of the most crucial safety-systems of a living being. It has limit switches, heat sensors, current sensors, strain gauges and alarm signals in all the right places to prevent it from destroying its many motors and joints. It has enormous "funny bones"--motors sticking out from its elbows in a risky way. These will be protected from harm not by being shielded in heavy armor, but by being equipped with patches of exquisitely sensitive piezo-electric membrane "skin" which will trigger alarms when they make contact with anything. The goal is that Cog will quickly "learn" to keep its funny bones from being bumped--if Cog cannot learn this in short order, it will have to have this high-priority policy hard-wired in. The same sensitive membranes will be used on its fingertips and elsewhere, and, like human tactile nerves, the "meaning" of the signals sent along the attached wires will depend more on what the central control system "makes of them" than on their "intrinsic"

characteristics. A gentle touch, signalling sought-for contact with an object to be grasped, will not differ, as an information packet, from a sharp pain, signalling a need for rapid countermeasures. It all depends on what the central system is designed to do with the packet, and this design is itself indefinitely revisable--something that can be adjusted either by Cog's own experience or by the tinkering of Cog's artificers.

One of its most interesting "innate" endowments will be software for visual face recognition. Faces will "pop out" from the background of other objects as items of special interest to Cog. It will further be innately designed to "want" to keep it's "mother's" face in view, and to work hard to keep "mother" from turning away. The role of mother has not yet been cast, but several of the graduate students have been tentatively tapped for this role. Unlike a human infant, of course, there is no reason why Cog can't have a whole team of mothers, each of whom is innately distinguished by Cog as a face to please if possible. Clearly, even if Cog really does have a *Lebenswelt*, it will not be the same as *ours*.

Decisions have not yet been reached about many of the candidates for hard-wiring or innate features. Anything that can learn must be initially equipped with a great deal of unlearned design. That is no longer an issue; no *tabula rasa* could ever be impressed with knowledge from experience. But it is also not much of an issue which features ought to be innately fixed, for there is a convenient trade-off. I haven't mentioned yet that Cog will actually be a multi-generational series of ever improved models (if all goes well!), but of course that is the way any complex artifact gets designed. Any feature that is not innately fixed at the outset, but does get itself designed into Cog's control system through learning, can then be lifted whole into Cog-II, as a new bit of innate endowment designed by Cog itself--or rather by Cog's history of interactions with its environment. So even in cases in which we have the best of reasons for thinking that human infants actually come innately equipped with pre-designed gear, we may choose to try to get Cog to learn the design in question, rather than be born with it. In some instances, this is laziness or opportunism--we don't really know what might work well, but maybe Cog can train itself up. This insouciance about the putative nature/nurture boundary is already a familiar attitude among neural net modelers, of course. Although Cog is not specifically intended to demonstrate any particular neural net thesis, it should come as no surprise that Cog's nervous system is a massively parallel architecture capable of simultaneously training up an indefinite number of special-purpose networks or circuits, under various regimes.

How plausible is the hope that Cog can retrace the steps of millions of years of evolution in a few months or years of laboratory exploration? Notice first that what I have just described is a variety of Lamarckian inheritance that no organic lineage has been able to avail itself of. The acquired design innovations of Cog-I can be immediately transferred to Cog-II, a speed-up of evolution of tremendous, if incalculable, magnitude. Moreover, if you bear in mind that, unlike the natural case, there will be a team of overseers ready to make patches whenever obvious shortcomings reveal themselves, and to jog the systems out of ruts whenever they enter them, it is not so outrageous a hope, in our opinion. But then, we are all rather outrageous people.

One talent that we have hopes of teaching to Cog is a rudimentary capacity for human language. And here we run into the fabled innate language organ or Language Acquisition Device made famous by Noam Chomsky. Is there going to be an attempt to build an innate LAD for our Cog? No. We are going to try to get Cog to build language the hard way, the way our ancestors must have done, over thousands of generations. Cog has ears (four, because it's easier to get good localization with four microphones than with carefully shaped ears like ours!) and some special-purpose signal-analyzing software is being developed to give Cog a fairly good chance of discriminating human speech sounds, and probably the capacity to distinguish different human voices. Cog will also have to have speech synthesis hardware and software, of course, but decisions have not yet been reached about the details. It is important to have Cog as well-equipped as possible for rich and natural interactions with human beings, for the team intends to take advantage of as much free labor as it can. Untrained people ought to be able to spend time--hours if they like, and we rather hope they do--trying to get Cog to learn this or that. Growing into an adult is a long,

time-consuming business, and Cog--and the team that is building Cog--will need all the help it can get.

Obviously this will not work unless the team manages somehow to give Cog a motivational structure that can be at least dimly recognized, responded to, and exploited by naive observers. In short, Cog should be as human as possible in its wants and fears, likes and dislikes. If those anthropomorphic terms strike you as unwarranted, put them in scare-quotes or drop them altogether and replace them with tedious neologisms of your own choosing: Cog, you may prefer to say, must have *goal-registrations* and *preference-functions* that map in rough isomorphism to human desires. This is so for many reasons, of course. Cog won't work at all unless it has its act together in a daunting number of different regards. It must somehow delight in learning, abhor error, strive for novelty, recognize progress. It must be vigilant in some regards, curious in others, and deeply unwilling to engage in self-destructive activity. While we are at it, we might as well try to make it crave human praise and company, and even exhibit a sense of humor.

Let me switch abruptly from this heavily anthropomorphic language to a brief description of Cog's initial endowment of information-processing hardware. The computer-complex that has been built to serve as the development platform for Cog's artificial nervous system consists of four backplanes, each with 16 nodes; each node is basically a Mac-II computer--a 68332 processor with a megabyte of RAM. In other words, you can think of Cog's brain as roughly equivalent to sixty-four Mac-IIs yoked in a custom parallel architecture. Each node is itself a multiprocessor, and they all run a special version of parallel Lisp developed by Rodney Brooks, and called, simply, L. Each node has an interpreter for L in its ROM, so it can execute L files independently of every other node.

Each node has 6 assignable input-output ports, in addition to the possibility of separate i-o (input-output) to the motor boards directly controlling the various joints, as well as the all-important i-o to the experimenters' monitoring and control system, the Front End Processor or FEP (via another unit known as the Interfep). On a bank of separate monitors, one can see the current image in each camera (two foveas, two parafoveas), the activity in each of the many different visual processing areas, or the activities of any other nodes. Cog is thus equipped at birth with the equivalent of chronically implanted electrodes for each of its neurons; all its activities can be monitored in real time, recorded and debugged. The FEP is itself a Macintosh computer in more conventional packaging. At startup, each node is awakened by a FEP call that commands it to load its appropriate files of L from a file server. These files configure it for whatever tasks it has currently been designed to execute. Thus the underlying hardware machine can be turned into any of a host of different virtual machines, thanks to the capacity of each node to run its current program. The nodes do not make further use of disk memory, however, during normal operation. They keep their transient memories locally, in their individual megabytes of RAM. In other words, Cog stores both its genetic endowment (the virtual machine) and its long term memory on disk when it is shut down, but when it is powered on, it first configures itself and then stores all its short term memory distributed one way or another among its 64 nodes.

The space of possible virtual machines made available and readily explorable by this underlying architecture is huge, of course, and it covers a volume in the space of all computations that has not yet been seriously explored by artificial intelligence researchers. Moreover, the space of possibilities it represents is manifestly much more realistic as a space to build brains in than is the space heretofore explored, either by the largely serial architectures of GOFAI ("Good Old Fashioned AI", Haugeland, 1985), or by parallel architectures simulated by serial machines. Nevertheless, it is arguable that every one of the possible virtual machines executable by Cog is minute in comparison to a real human brain. In short, Cog has a tiny brain. There is a big wager being made: the parallelism made possible by this arrangement will be sufficient to provide real-time control of importantly humanoid activities occurring on a human time scale. If this proves to be too optimistic by as little as an order of magnitude, the whole project will be forlorn, for the motivating insight for the project is that by confronting and solving *actual, real time* problems of self-protection, hand-eye coordination, and interaction with other animate beings, Cog's artificers will discover the *sufficient*

conditions for higher cognitive functions in general--and maybe even for a variety of consciousness that would satisfy the skeptics.

It is important to recognize that although the theoretical importance of having a body has been appreciated ever since Alan Turing (1950) drew specific attention to it in his classic paper, "Computing Machines and Intelligence," within the field of Artificial Intelligence there has long been a contrary opinion that robotics is largely a waste of time, money and effort. According to this view, whatever deep principles of organization make cognition possible can be as readily discovered in the more abstract realm of pure simulation, at a fraction of the cost. In many fields, this thrifty attitude has proven to be uncontroversial wisdom. No economists have asked for the funds to implement their computer models of markets and industries in tiny robotic Wall Streets or Detroits, and civil engineers have largely replaced their scale models of bridges and tunnels with computer models that can do a better job of simulating all the relevant conditions of load, stress and strain. Closer to home, simulations of ingeniously oversimplified imaginary organisms foraging in imaginary environments, avoiding imaginary predators and differentially producing imaginary offspring are yielding important insights into the mechanisms of evolution and ecology in the new field of Artificial Life. So it is something of a surprise to find this AI group conceding, in effect, that there is indeed something to the skeptics' claim (e.g., Dreyfus and Dreyfus, 1986) that genuine embodiment in a real world is crucial to consciousness. Not, I hasten to add, because genuine embodiment provides some special vital juice that mere virtual-world simulations cannot secrete, but for the more practical reason--or hunch--that unless you saddle yourself with all the problems of making a concrete agent take care of itself in the real world, you will tend to overlook, underestimate, or misconstrue the deepest problems of design.

Besides, as I have already noted, there is the hope that Cog will be able to design itself in large measure, learning from infancy, and building its own representation of its world in the terms that it innately understands. Nobody doubts that any agent capable of interacting intelligently with a human being on human terms must have access to literally millions if not billions of logically independent items of world knowledge. Either these must be hand-coded individually by human programmers--a tactic being pursued, notoriously, by Douglas Lenat and his CYC team in Dallas--or some way must be found for the artificial agent to learn its world knowledge from (real) interactions with the (real) world. The potential virtues of this shortcut have long been recognized within AI circles (e.g., Waltz, 1988). The unanswered question is whether taking on the task of solving the grubby details of real-world robotics will actually permit one to finesse the task of hand-coding the world knowledge. Brooks, Stein and their team--myself included--are gambling that it will.

At this stage of the project, most of the problems being addressed would never arise in the realm of pure, disembodied AI. How many separate motors might be used for controlling each hand? They will have to be mounted somehow on the forearms. Will there then be room to mount the motor boards directly on the arms, close to the joints they control, or would they get in the way? How much cabling can each arm carry before weariness or clumsiness overcome it? The arm joints have been built to be compliant--springy, like your own joints. This means that if Cog wants to do some fine-fingered manipulation, it will have to learn to "burn" some of the degrees of freedom in its arm motion by temporarily bracing its elbows or wrists on a table or other convenient landmark, just as you would do. Such compliance is typical of the mixed bag of opportunities and problems created by real robotics. Another is the need for self-calibration or re-calibration in the eyes. If Cog's eyes jiggle away from their preset aim, thanks to the wear and tear of all that sudden saccading, there must be ways for Cog to compensate, short of trying continually to adjust its camera-eyes with its fingers. Software designed to tolerate this probable sloppiness in the first place may well be more robust and versatile in many other ways than software designed to work in a more "perfect" world.

Earlier I mentioned a reason for using artificial muscles, not motors, to control a robot's joints, and the example was not imaginary. Brooks is concerned that the sheer noise of Cog's skeletal activities may seriously interfere with the attempt to give Cog humanoid hearing. There is research underway at the AI Lab

to develop synthetic electro-mechanical muscle tissues, which would operate silently as well as being more compact, but this will not be available for early incarnations of Cog. For an entirely different reason, thought is being given to the option of designing Cog's visual control software *as if* its eyes were moved by muscles, not motors, building in a software interface that amounts to giving Cog a set of *virtual* eye-muscles. Why might this extra complication in the interface be wise? Because the "opponent-process" control system exemplified by eye-muscle controls is apparently a deep and ubiquitous feature of nervous systems, involved in control of attention generally and disrupted in such pathologies as unilateral neglect. If we are going to have such competitive systems at higher levels of control, it might be wise to build them in "all the way down," concealing the final translation into electric-motor-talk as part of the backstage implementation, not the model.

Other practicalities are more obvious, or at least more immediately evocative to the uninitiated. Three huge red "emergency kill" buttons have already been provided in Cog's environment, to ensure that if Cog happens to engage in some activity that could injure or endanger a human interactor (or itself), there is a way of getting it to stop. But what is the appropriate response for Cog to make to the KILL button? If power to Cog's motors is suddenly shut off, Cog will slump, and its arms will crash down on whatever is below them. Is this what we want to happen? Do we want Cog to drop whatever it is holding? What should "Stop!" *mean* to Cog? This is a real issue about which there is not yet any consensus.

There are many more details of the current and anticipated design of Cog that are of more than passing interest to those in the field, but on this occasion, I want to use the little remaining time to address some overriding questions that have been much debated by philosophers, and that receive a ready treatment in the environment of thought made possible by Cog. In other words, let's consider Cog merely as a prosthetic aid to philosophical thought-experiments, a modest but by no means negligible role for Cog to play.

3. Some Philosophical Considerations

A recent criticism of "strong AI" that has received quite a bit of attention is the so-called problem of "symbol grounding" (Harnad, 1990). It is all very well for large AI programs to have data structures that *purport* to refer to Chicago, milk, or the person to whom I am now talking, but such imaginary reference is not the same as real reference, according to this line of criticism. These internal "symbols" are not properly "grounded" in the world, and the problems thereby eschewed by pure, non-robotic, AI are not trivial or peripheral. As one who discussed, and ultimately dismissed, a version of this problem many years ago (Dennett, 1969, p.182ff), I would not want to be interpreted as now abandoning my earlier view. I submit that Cog moots the problem of symbol grounding, without having to settle its status as a criticism of "strong AI". Anything in Cog that might be a candidate for symbolhood will automatically be "grounded" in Cog's real predicament, as surely as its counterpart in any child, so the issue doesn't arise, except as a practical problem for the Cog team, to be solved or not, as fortune dictates. If the day ever comes for Cog to comment to anybody about Chicago, the question of whether Cog is in any position to do so will arise for exactly the same reasons, and be resolvable on the same considerations, as the parallel question about the reference of the word "Chicago" in the idiolect of a young child.

Another claim that has often been advanced, most carefully by Haugeland (1985), is that nothing could properly "matter" to an artificial intelligence, and mattering (it is claimed) is crucial to consciousness. Haugeland restricted his claim to traditional GOFAI systems, and left robots out of consideration. Would he concede that something could matter to Cog? The question, presumably, is how seriously to weigh the import of the quite deliberate decision by Cog's creators to make Cog as much as possible responsible for its own welfare. Cog will be equipped with some "innate" but not at all arbitrary preferences, and hence provided of necessity with the concomitant capacity to be "bothered" by the thwarting of those preferences, and "pleased" by the furthering of the ends it was innately designed to seek. Some may want to retort: "This is not *real* pleasure or pain, but merely a simulacrum." Perhaps, but on what grounds will they defend this

claim? Cog may be said to have quite crude, simplistic, one-dimensional pleasure and pain, cartoon pleasure and pain if you like, but then the same might also be said of the pleasure and pain of simpler organisms--clams or houseflies, for instance. Most, if not all, of the burden of proof is shifted by Cog, in my estimation. The reasons for saying that something *does* matter to Cog are not arbitrary; they are exactly parallel to the reasons we give for saying that things matter to us and to other creatures. Since we have cut off the dubious retreats to vitalism or origin chauvinism, it will be interesting to see if the skeptics have any good reasons for declaring Cog's pains and pleasures not to matter--at least to it, and for that very reason, to us as well. It will come as no surprise, I hope, that more than a few participants in the Cog project are already musing about what obligations they might come to have to Cog, over and above their obligations to the Cog team.

Finally, J.R. Lucas (1994) has raised the claim that if a robot were really conscious, we would have to be prepared to believe it about its own internal states. I would like to close by pointing out that this is a rather likely reality in the case of Cog. Although equipped with an optimal suite of monitoring devices that will reveal the details of its inner workings to the observing team, Cog's own pronouncements could very well come to be a more trustworthy and informative source of information on what was really going on inside it. The information visible on the banks of monitors, or gathered by the gigabyte on hard disks, will be at the outset almost as hard to interpret, even by Cog's own designers, as the information obtainable by such "third-person" methods as MRI and CT scanning in the neurosciences. As the observers refine their models, and their understanding of their models, their authority as interpreters of the data may grow, but it may also suffer eclipse. Especially since Cog will be designed from the outset to redesign itself as much as possible, there is a high probability that the designers will simply lose the standard hegemony of the artificer ("I made it, so I know what it is supposed to do, and what it is doing now!"). Into this epistemological vacuum Cog may very well thrust itself. In fact, I would gladly defend the conditional prediction: *if* Cog develops to the point where it can conduct what appear to be robust and well-controlled conversations in something like a natural language, it will certainly be in a position to rival its own monitors (and the theorists who interpret them) as a source of knowledge about what it is doing and feeling, and why.

References

Dennett, Daniel C., 1969, *Content and Consciousness*, London: Routledge & Kegan Paul.

Dennett, Daniel C., 1987, "Fast Thinking," in Dennett, *The Intentional Stance*, Cambridge, MA: MIT Press, pp. 323-37.

Dennett, Daniel C., 1993a, review of John Searle, *The Redisovery of the Mind*, in *J.Phil*. **90**, pp.193-205.

Dennett, Daniel C., 1993b, "Caveat Emptor," *Consciousness and Cognition*, **2**, pp.48-57.

Dreyfus, Hubert & Dreyfus, Stuart, 1986, *Mind Over Machine*, New York: MacMillan.

Harnad, Stevan, 1990, "The Symbol Grounding Problem," *Physica D*, **42**, pp.335-46.

Haugeland, John, 1985, *Artificial Intelligence: The Very Idea*, Cambridge MA: MIT Press.

Lucas, J. R., 1994, [presentation to the Royal Society, Conference on Artificial Intelligence, April 14, 1994.

Mangan, Bruce, "Dennett, Consciousness, and the Sorrows of Functionalism," *Consciousness and Cognition*, **2**, pp-1-17.

Searle, John, 1992, *The Redisovery of the Mind*, Cambridge, MA: MIT Press.

Turing, Alan, 1950, "Computing Machinery and Intelligence," *Mind*, **59**, pp.433-60.

Waltz, David, 1988, "The Prospects for Building Truly Intelligent Machines," *Daedalus*, **117**, pp.191-222.