Appendix: Machine Consciousness

Artificial Intelligence is, very crudely, the science of getting machines to perform jobs that normally require intelligence and judgment. Researchers at any number of AI labs have designed machines that prove mathematical theorems, play chess, sort mail, guide missiles, assemble auto engines, diagnose illnesses, read stories and other written texts, and converse with people in a rudimentary way. This is, we might say, intelligent behavior.

But what is this "intelligence"? As a first pass, I suggest that intelligence of the sort I am talking about is a kind of flexibility, a responsiveness to contingencies. A dull or stupid machine must have just the right kind of raw materials presented to it in just the right way, or it is useless: the electric can opener must have an appropriately sized can fixed under its drive wheel *just so*, in order to operate at all. Humans (most of us, anyway) are not like that. We deal with the unforeseen. We take what comes and make the best of it, even though we may have had no idea what it would be. We play the ball from whatever lie we are given, and at whatever angle to the green; we read and understand texts we have never seen before; we find our way back to Chapel Hill after getting totally lost in downtown Durham (or downtown Washington, D.C., or downtown Lima, Peru).

Our pursuit of our goals is guided while in progress by our ongoing perception and handling of interim developments. Moreover, we can pursue any number of different goals at the same time, and balance them against each other. We are sensitive to contingencies, both external and internal, that have a very complex and unsystematic structure.

It is almost irresistible to speak of *information* here, even if the term were not as trendy as it is. An intelligent creature, I want to say, is an *information-sensitive* creature, one that not only *registers* information through receptors such as sense-organs but somehow stores and manages and finally uses that information. Higher animals are intelligent beings in this sense, and so are we, even though virtually nothing is known about how we organize or manage the vast, seething

profusion of information that comes our way. And there is one sort of machine that is information-sensitive also: the digital computer. A computer *is* a machine specifically designed to be fed complexes of information, to store them, manage them, and produce appropriate theoretical or practical conclusions on demand. Thus, if artificial intelligence is what one is looking for, it is no accident that one looks to the computer.

Yet a computer has two limitations in common with machines of less elite and grandiose sorts, both of them already signaled in the characterization I have just given. First, a (present-day) computer must be *fed* information, and the choice of what information to feed and in what form is up to a human programmer or operator. (For that matter, a present-day computer must be plugged into an electrical outlet and have its switch turned to ON, but this is a very minor contingency given the availability of nuclear power packs.) Second, the *appropriateness* and effectiveness of a computer's output depends entirely on what the programmer or operator had in mind and goes on to make of it. A computer has intelligence in the sense I have defined, but has no judgment, since it has no goals and purposes of its own and no internal sense of appropriateness, relevance, or proportion.

For essentially these reasons—that computers are intelligent in my minimal sense, and that they are nevertheless limited in the two ways I have mentioned—AI theorists, philosophers, and intelligent laymen have inevitably compared computers to human minds, but at the same time debated both technical and philosophical questions raised by this comparison. The questions break down into three main groups or types: (A) Questions of the form "Will a computer ever be able to do *X*?" where *X* is something that intelligent humans can do. (B) Questions of the form "Given that a computer can or could do *X*, have we any reason to think that it does *X* in the same way that humans do *X*?" (C) Questions of the form "Given that some futuristic supercomputer were able to do *X*, *Y*, *Z*, . . . , for some arbitrarily large range and variety of human activities, would that show that the computer had property *P*?" where *P* is some feature held to be centrally, vitally characteristic of human minds, such as thought, consciousness, feeling, sensation, emotion, creativity, or freedom of the will.

Questions of type A are empirical questions and cannot be settled without decades, perhaps centuries, of further research—compare ancient and medieval speculations on the question of whether a machine could ever fly. Questions of type B are brutely empirical too, and their answers are unavailable to AI researchers *per se*, lying squarely in the domain of cognitive psychology, a science or alleged

science barely into its infancy. Questions of type C are philosophical and conceptual, and so I shall essay to answer them all at one stroke.

Let us begin by supposing that all questions of types A and B have been settled affirmatively—that one day we might be confronted by a much-improved version of Hal, the soft-spoken computer in Kubrick's *2001* (younger readers may substitute Star Wars' C3PO or whatever subsequent cinematic robot is the most lovable). Let us call this more versatile machine "Harry."[2] Harry (let us say) is humanoid in form—he is a miracle of miniaturization and has lifelike plastic skin—and he can converse intelligently on all sorts of subjects, play golf *and* the viola, write passable poetry, control his occasional nervousness pretty well, make love, prove mathematical theorems (of course), show envy when outdone, throw gin bottles at annoying children, etc., etc. We may suppose he fools people into thinking he is human. Now the question is, is Harry really a *person?* Does he have thoughts, feelings, and so on? Is he actually conscious, or is he just a mindless walking hardware store whose movements are astoundingly *like* those of a person?[3]

Plainly his acquaintances would tend from the first to see him as a person, even if they were aware of his dubious antecedents. I think it is a plain psychological fact, if nothing more, that we could not help treating him as a person, unless we resolutely made up our minds, on principle, not to give him the time of day. But how could we really tell that he is conscious?

Well, how do we really tell that any humanoid creature is conscious? How do you tell that I am conscious, and how do I tell that you are? Surely we tell, and decisively, on the basis of our standard behavioral tests for mental states, to revert to a theme of chapter 3: We know that a human being has such-and-such mental states when it behaves, to speak very generally, in the ways we take to be appropriate to organisms that are in those states. (The point is of course an epistemological one only, no metaphysical implications intended or tolerated.) We know for practical purposes that a creature has a mind when it fulfills all the right criteria. And by hypothesis, Harry fulfills all our behavioral criteria with a vengeance; moreover, he does so *in the right way* (cf. questions of type B): the processing that stands causally behind his behavior is just like ours. It follows that we are at least *prima facie* justified in believing him to be conscious.

We have not *proved* that he is conscious, of course—any more than you have proved that I am conscious. An organism's merely behaving in a certain way is no logical guarantee of sentience; from my point of view it is at least imaginable, a bare logical possibility, that my wife, my daughter, and my chairman are not conscious, even though I have

excellent, overwhelming behavioral reason to think that they are. But for that matter, our "standard behavioral tests" for mental states yield practical or moral certainty only so long as the situation is not palpably extraordinary or bizarre. A human chauvinist—in this case, someone who denies that Harry has thoughts and feelings, joys and sorrows—thinks precisely that Harry is as bizarre as they come. But *what is bizarre about him?* There are quite a few chauvinist answers to this, but what they boil down to, and given our hypothesized facts all they could boil down to, are two differences between Harry and ourselves: his *origin* (a laboratory is not a proper mother), and the *chemical composition of his anatomy,* if his creator has used silicon instead of carbon, for example. To exclude him from our community for either or both of *those* reasons seems to me to be a clear case of racial or ethnic prejudice (literally) and nothing more. I see no obvious way in which either a creature's origin or its subneuroanatomical chemical composition should matter to its psychological processes or any aspect of its mentality.

My argument can be reinforced by a thought-experiment, in the spirit of chapters 3 and 5: Imagine that we take a normal human being, Henrietta, and begin gradually replacing parts of her with synthetic materials—first a few prosthetic limbs, then a few synthetic arteries, then some neural fibers, and so forth. Suppose that the surgeons who perform the successive operations (particularly the neurosurgeons) are so clever and skillful that Henrietta survives in fine style: her intelligence, personality, perceptual acuity, poetic abilities, etc., remain just as they were before. But after the replacement process has eventually gone on to completion, Henrietta will have become an artifact—at least, her body will then be nothing but a collection of artifacts. Did she lose consciousness at some point during the sequence of operations, despite her continuing to behave and respond normally? When? It is hard to imagine that there is some privileged portion of the human nervous system that is for some reason indispensable, even though kidneys, lungs, heart, and any given bit of brain could in principle be replaced by a prosthesis (for *what* reason?); and it is also hard to imagine that there is some *pro*portion of the nervous system such that removal of more than that proportion causes loss of consciousness or sentience despite perfect maintenance of all intelligent capacities.

If this quick but totally compelling defense of Harry and Henrietta's personhood is correct, then the two, and their ilk, will have not only mental lives like ours, but *moral* lives like ours, and moral rights and privileges accordingly. Just as origin and physical constitution fail to affect psychological personhood, if a creature's internal organization

is sufficiently like ours, so do they fail to affect moral personhood. We do not discriminate against a person who has a wooden leg, or a mechanical kidney, or a nuclear heart regulator; no more should we deny any human or civil right to Harry or Henrietta on grounds of their origin or physical makeup, which they cannot help.

But this happy egalitarianism raises a more immediate question: *In real life,* we shall soon be faced with medium-grade machines, which have some intelligence and are not "mere" machines like refrigerators or typewriters but which fall far short of flawless human simulators like Harry. For AI researchers may well build machines that will appear to have some familiar mental capacities but not others. The most obvious example is that of a sensor or perceptron, which picks up information from its immediate environment, records it, and stores it in memory for future printout. (We already have at least crude machines of this kind. When they become versatile and sophisticated enough, it will be quite natural to say that they see or hear and that they remember.) But the possibility of "specialist" machines of this kind raises an unforeseen contingency: There is an enormous and many-dimensional range of possible beings in between our current "mere" machines and our fully developed, flawless human simulators; we have not even begun to think of all the infinitely possible variations on this theme. And once we do begin to think of these hard cases, we will be at a loss as to where to draw the "personhood" line between them. How complex, eclectic, and impressive must a machine be, and in what respects, before we award it the accolade of personhood and/or of consciousness? There is, to say the least, no clear answer to be had *a priori,* Descartes' notorious view of animals to the contrary notwithstanding.

This typical philosophical question would be no more than an amusing bonbon, were it not for the attending moral conundrum: What moral rights would an intermediate or marginally intelligent machine have? Adolescent machines of this sort will confront us much sooner than will any good human simulators, for they are easier to design and construct; more to the moral point, they will be designed mainly as *labor-saving devices,* as servants who will work for free, and servants of this kind are (literally) made to be exploited. If they are intelligent to any degree, we should have qualms in proportion.

I suggest that this moral problem, which may become a real and pressing one, is parallel to the current debate over animal rights. Luckily I have never wanted to cook and eat my Compaq Portable.

Suppose I am right about the irrelevance of biochemical constitution to psychology; and suppose I was also right about the coalescing

of the notions *computation, information, intelligence.* Then our mentalized theory of computation suggests in turn a computational theory of mentality, and a computational picture of the place of human beings in the world. In fact, philosophy aside, that picture has already begun to get a grip on people's thinking—as witness the filtering down of computer jargon into contemporary casual speech—and that grip is not going to loosen. Computer science is the defining technology of our time, and in this sense the computer is the natural cultural successor to the steam engine, the clock, the spindle, and the potter's wheel.[4] Predictably, an articulate computational theory of the mind has also gained credence among professional psychologists and philosophers.[5] I have been trying to support it here and elsewhere; I shall say no more about it for now, save to note again its near-indispensability in accounting for intentionality (noted), and to address the ubiquitous question of computer creativity and freedom:

Soft Determinism or Libertarianism may be true of humans. But many people have far more rigidly deterministic intuitions about computers. Computers, after all, (let us all say it together:) "only do what they are told/programmed to do"; they have no spontaneity and no freedom of choice. But human beings choose all the time, and the ensuing states of the world often depend entirely on these choices.[6] Thus the "computer analogy" supposedly fails.

The alleged failure of course depends on what we think freedom really is. As a Soft Determinist, I think that to have freedom of choice in acting is (roughly) for one's action to proceed out of one's own desires, deliberation, will, and intention, rather than being compelled or coerced by external forces regardless of my desires or will. As before, free actions are not *uncaused* actions. My free actions are those that *I* cause, i.e., that are caused by my own mental processes rather than by something pressing on me from the outside. I have argued in chapter 9 that I am free in that my beliefs, desires, deliberations, and intentions are all functional or computational states and processes within me that do interact in characteristic ways to produce my behavior. Note now that the same response vindicates our skilled human-simulating machines from the charge of puppethood. The word "robot" is often used as a veritable synonym for "puppet," so it may seem that Harry and Henrietta are paradigm cases of *un*free mechanisms that "only do what they are programmed to do." This is a slander—for two reasons:

First, even an ordinary computer, let alone a fabulously sophisticated machine like Harry, is in a way unpredictable. You are at its mercy. You *think* you know what it is going to do; you know what it should do, what it is supposed to do, but there is no guarantee—and

it may do something *awful* or at any rate something that you could not have predicted and could not figure out if you tried with both hands. This practical sort of unpredictability would be multiplied a thousandfold in the case of a machine as complex as the human brain, and it is notably characteristic of *people.*

The unpredictability has several sources. (i) Plain old physical defects, as when Harry's circuits have been damaged by trauma, stress, heat, or the like. (ii) Bugs in one or more of his programs. (I have heard that once upon a time, somewhere, a program was written that had not a single bug in it, but this is probably an urban folk tale.) (iii) Randomizers, quantum-driven or otherwise; elements of Harry's behavior may be *genuinely,* physically random. (iv) Learning and analogy mechanisms; if Harry is equipped with these, as he inevitably would be, then his behavior-patterns will be modified in response to his experiential input from the world, which would be neither controlled nor even observed by us. *We don't know where he's been.* (v) The relativity of reliability to goal-description. This last needs a bit of explanation.

People often say things like, "A computer just crunches binary numbers; provided it isn't broken, it just chugs on mindlessly through whatever flipflop settings are predetermined by its electronic makeup." But such remarks ignore the multileveled character of real computer programming. At any given time, as we have noted in chapter 4, a computer is running *each of any number of* programs, depending on how it is described and on the level of functional organization that interests us. True, it is always crunching binary numbers, but in crunching them it is also doing any number of more esoteric things. And (more to the point) what counts as a mindless, algorithmic procedure at a very low level of organization may constitute, at a higher level, a hazardous do-or-die heuristic that might either succeed brilliantly or (more likely) fail and leave its objective unfulfilled.

As a second defense, remember that Harry too has beliefs, desires, and intentions (provided my original argument is sound). If this is so, then his behavior normally proceeds out of his own mental processes rather than being externally compelled; and so he satisfies the definition of freedom-of-action formulated above. In most cases it will be appropriate to say that Harry could have done other than what he did do (but in fact chose after some ratiocination to do what he did, instead). Harry acts in the same sense as that in which we act, though one might continue to quarrel over what sense that is.

Probably the most popular remaining reason for doubt about machine consciousness has to do with the raw qualitative character of

experience. Could a mere bloodless runner-of-programs have states that *feel to it* in any of the various dramatic ways in which our mental states feel to us?

The latter question is usually asked rhetorically, expecting a resounding answer "NO!!" But I do not hear it rhetorically, for I do not see why the negative answer is supposed to be at all obvious, even for machines as opposed to biologic humans. Of course there is an incongruity *from our human point of view* between human feeling and printed circuitry or silicon pathways; that is to be expected, since we are considering those high-tech items from an external, third-person perspective and at the same time comparing them to our own first-person feels. But argumentatively, that *Gestalt* phenomenon counts for no more in the present case than it did in that of human consciousness, viz., for nothing, especially if my original argument about Harry was successful in showing that biochemical constitution is irrelevant to psychology. What matters to mentality is not the stuff of which one is made, but the complex way in which that stuff is organized.[7] If after years of close friendship we were to open Harry up and find that he is stuffed with microelectronic gadgets instead of protoplasm, we would be taken aback—no question. But our *Gestalt* clash on the occasion would do nothing *at all* to show that Harry does not have his own rich inner qualitative life. If an objector wants to insist that computation alone cannot provide consciousness with its qualitative character, the objector will have to take the initiative and come up with a further, substantive argument to show why not.[8] We have already seen that such arguments have failed wretchedly for the case of humans; I see no reason to suspect that they would work any better for the case of robots. We must await further developments. But at the present stage of inquiry I see no compelling feel-based objection to the hypothesis of machine consciousness.

part I have undertaken in this book. If I have failed, I would like to be *shown why* (or, of course, presented with some new antimaterialist argument). To engage in further muttering and posturing would be idle.

2. Perhaps it is time for a brisk catalogue, all in one spot, of the different "qualia"-based objections we have encountered in this book, which are all the different ones I myself have encountered anywhere: (i) Early critics of the Identity Theory invoked qualia in posing Leibniz's-Law objections (see Lycan, 1972, and the references therein). (ii) Others focused on our seemingly immediate access to qualia (e.g., Baier, 1962). (iii) As was discussed in chapter 2, Saul Kripke's rejection of physicalism is based on an essentialist thesis involving qualia. (iv) Still other philosophers have pursued the sort of counterexample technique discussed in chapter 3. (v) Nagel, and Keith Gunderson (1970, 1974) before him, have worried over first-person/third-person asymmetries and the perspectival nature or point-of-viewiness of consciousness. (vi) As I have read them in chapter 7, Nagel and Frank Jackson also call our attention to what he thinks is a funny kind of *fact* that has no place in physical science. (vii) Jackson earlier argued for the existence of little colored nonphysical sense-data in (or near) the head, and we have seen in chapter 8 that the appeal to phenomenal individuals is a powerful antimaterialist force, especially when subtly introduced by way of the Banana Peel. Finally, as distinct from all these concerns, (viii) Sellars has stressed the grainlessness or *homogeneity* of sensory qualia, and maintained that that homogeneity is what prevents our dissolving qualia peacefully into a Democritean picture of the universe. If there are still more different "qualia" arguments, I have failed to discern them.

Appendix

1. The material in this appendix was first presented as part of the John Ingram Forry Lecture at Amherst College, in 1985. I am very grateful to Jay Garfield and to Lee Bowie for their penetrating formal commentaries on that occasion, which I shall be answering in the (eventually to be) published proceedings of the event.
2. Harry has appeared before, in Lycan (1985). The next four paragraphs are lifted almost *verbatim* from that article.
3. It is interesting that children seem instinctively to reject the hypothesis of machine consciousness, usually on the grounds that computers are not alive. (One day when my daughter Jane was three years old, we were fooling with some piece of software or other, and I quite unreflectively remarked "It thinks you want it to [do such-and-such]." She did an enormous take, and then replied, "Computers can't think!—Is that 'just an expression'??")
4. I borrow the term "defining technology," and the examples, from Jay Bolter (1984).
5. The computational picture of mentality is by no means new. For one thing, the idea of mechanical intelligence goes back to the seventeenth century at least, long before Charles Babbage's celebrated Analytical Engine. And the computer model of the mind received a decisive boost from the McCullough-Pitts model of the neuron (1947), according to which a neuron is nothing but a little on-off device, that either *fires* or does not. If a brain is just an organized collection of neurons, and a neuron is just an on-off switch, it follows *straightway* that a brain is a digital computer and anything interesting that it does is a computation over binary formulas. Thus a human being is not only a featherless biped, a rational animal, and the only creature on earth that laughs, but the only computing machine on earth that is made by unskilled labor.

   The McCullough-Pitts model is no longer current (no pun intended): neurons are

now known to be very complicated little agents, not mere on-off switches. But the computational picture of mentality still receives strong encouragement from other quarters. It has two separate philosophical motivations, in particular, the first of which I have already noted: It exploits and explains the coalescence of the notions of computation, information, and intelligence. The computer is the only thing in the world that displays potential intelligence *and* whose workings are well understood. It is the only answer we currently know to the question: By what means *could* Mother Nature have crafted an intelligent being (in our sense of responsiveness to contingencies) out of nothing but a large bunch of individually insensate biological cells? To deny that there may be other answers would be presumptuous at best, and there are plenty of human capacities that do not seem to admit of computational simulation in any way at all—but anyone who manages to think up a genuinely distinct alternative to the digital-computer paradigm will have achieved a major conceptual breakthrough. For the foreseeable future, computation is our only model for intelligence.

Computationalism as a form of Homunctionalism also affords us a way of acknowledging our place as physical organisms amid the closed causal order we call Nature, without benefit of intervention by ghosts. (Actually I hear there are some physicists who speculate that quantum indeterminacies afford gaps in nature that are in principle permeable to Cartesian minds, and that immaterial egos do insert themselves into quantum gaps, thus taking over the role of hidden variables. But (i) it would have to be shown how such quantum phenomena could be combined and multiplied into macroscopic effects characteristic of intelligence, i.e., how the brain could act as a "quantum magnifier," and (ii) to avoid *ad-hoc*ness of the crassest sort, one would have to find *physical reason* to think that Cartesian intervention does occur, which task I take to be almost definitionally impossible.)

6. Of course, this re-emphasizes the question of human freedom: if humans are just wetware or liveware, are they not then essentially soft puppets? This in turn suggests—however speciously in light of the arguments made in chapter 9—that the computational view of people must therefore be drastically wrong.
7. Relatively speaking, of course; I am not encouraging Two-Levelism.
8. That mental acts do not *feel* digital is not an objection either. To infer from that fact that mental acts are not digital would be a clear case of what Armstrong (1968a) calls the "headless woman" fallacy.