METAPHILOSOPHY
Vol. 16, No. 4, October 1985
0026–1068 $2.00

FRANKENSTEIN'S CHILDREN: ARTIFICIAL INTELLIGENCE AND HUMAN VALUE*

DAN LLOYD

> "I am alone and miserable; man will not associate with me; but one as deformed and horrible as myself would not deny herself to me. My companion must be of the same species and have the same defects. This being you must create." – Mary Shelley, *Frankenstein* (New York: Bantam, 1967) p. 129.

Artificial Intelligence, according to one standard definition, is the field of computer science which seeks to program computers to perform tasks which, when performed by humans, require intelligence. Research in AI coincides with the history of the computer itself – Turing contributed to the conception of AI as early as 1950.[1] Since then, computer scientists have taken two broad approaches to work in AI: The first, the *cognitive simulation* approach regards AI as offering models of human cognition (Schank and Colby 1973; Newell 1973). The aim of simulation AI is to program computers not only to perform intelligent tasks, but to do them the way people do, that is, by employing similar cognitive processes. The second approach, the *engineering* approach, equally aims at the performance of intelligent tasks, but unlike cognitive simulation AI, does not intend to model human cognitive processes. Engineering AI takes advantage of *any* methods that efficiently accomplish the chosen tasks, whether the method is characteristically human or not (Nilsson 1971).

Both conceptions of AI raise serious ethical issues, issues which are rapidly becoming acute as AI leaves the laboratory for the marketplace. For example, engineering AI has begun to turn out practical "expert systems", large programs which encode and employ bodies of expertise in specific domains. Computers running these programs are currently capable of matching the performance of human experts in medical diagnosis, geological prediction, financial planning, and many other fields. In fact, "knowledge engineers" (as high-level expert systems programmers are called) can perform "knowledge extraction" with any expert, encoding his or her know-how in if-then rules to be

*Editor's footnote: This article was part of the essay competition on Computer Ethics.

[1] See Turing 1950. In that article, Turing proposed the classic "Turing test", according to which a machine would be deemed intelligent if it could pass for a human being communicating via teletype. The adequacy of the test has been subject to continued debate. See, for example, Chapter 4, "The Logic of Simulation", in Fodor 1968.

executed by the computer in unlimited novel situations. The widespread use of such programs immediately raises serious issues of *displacement*. In the AI engineers' world, expertise will no longer be a scarce human resource, but a commodity available – for a price – on floppy disks. The result is certain to be an extensive realignment of the professional workforce. Displacement also characterizes the effect of expert systems on *ethical responsibility*. To pose a simple example, suppose a hospital uses an expert system to make diagnoses, and in a particular case a diagnosis turns out to be fatally incorrect. If malpractice is indicated, who is liable? Consider everyone involved: The expert systems programmer, the experts who were consulted to write the program, the company marketing the program, the hospital which instituted the policy of reliance on the system's diagnoses, the hospital staffer who endorsed the diagnosis and initiated treatment. Tracing responsibility among them will pose significant moral and legal issues.

I mention these issues arising from engineering AI to point out their importance, but at the same time to suggest an analogy which makes them more tractable. The analogy is between expert systems and printed books. The advent of inexpensive printing following the invention of moveable type created similar displacements, especially surrounding the class of experts. Their expertise, harnessed by Renaissance "knowledge engineers" and their printing presses, suddenly became a ready commodity. End users, that is, readers, were now widely able to rely on the encapsulated advice of experts in every discipline without needing the expert to be present. And that reliance created similar realignments of ethical responsibility: If a book contained dangerous misinformation, then the same questions of liability arose as arise with contemporary expert systems.

My suggestion, then, is that the serious issues of engineering AI might turn out to be special cases of ethical issues that have been with us for centuries. Indeed, this is arguably the case with most new technologies, computer technology included. Computer theft and invasion of privacy, for example, are special cases of familiar ethical issues, and the task of the computer ethicists consists largely in properly aligning the new technology with preexisting conceptions of morality – a task which is neither obvious nor easy.

The issues surrounding engineering AI contrast with those of cognitive simulation AI, the main subject of this paper. Simulation AI raises an ethical issue with radically new dimensions, unparallelled by any other technology. This newness arises when one ponders the final extrapolation of a technology which simulates human cognitive processes in a fundamentally nonhuman medium. Simulation AI, taken to its limit, will produce *conscious machines*. I will argue that the issue of computer consciousness is not science fiction, arising from an overdose of movies and marketing hyperbole, but rather that conscious machines are conceptually and, in the end, technically feasible. Because such a machine is conscious, we stand in a unique ethical relation to it, complicated by the possibility that a machine consciousness does not share many of our human values and goals. My arguments concerning these

issues are presented in five distinct premises. On their basis I will conclude with a recommendation about the future of AI.

1. *Aritificial Minds Are Ultimately Feasible*

A common response to the proposal that a machine could embody a mind or become conscious is the confident intuition that "mere" matter can't be conscious, and that any appearance of consciousness in machines is just an ill-advised metaphorical extension of human animation to the inert. At the root of this intuition is dualism, sometimes a "closet dualism" which denies the central mind/body gulf of full-blown dualism while preserving some of the consequences of dualist metaphysics. Later I will argue that even if explicit Cartesian dualism should turn out to be true, that is irrelevant to the ethical issues discussed in this paper. At present, I hope to disable the intuition that machines *cannot* be conscious. I'll do that both on philosophical grounds and through a brief look at current progress in computers and psychology.

First, there is the oft-noted point that most of us *do* believe that matter can be conscious. Case in point: the matter between our ears. At first glance, our brains look no more promising than our livers as the seat of the mind, yet brains are indeed conscious. But between brain and digital computer lies a chasm, and it is here that the resistance to the conceptual possibility of conscious machines is probably deepest. Consider, then, the following thought experiment.

We begin with a *prosthetic neuron*. The prosthetic neuron takes advantage of the fact that nervous impulse conduction is partly an electrical process. The prosthesis is a minute microcircuit which responds, as do regular neurons, to the arrival of neurotransmitting chemicals at its artificial synaptic junctures. When it detects stimulation, the prosthesis sends out a signal to its axon processes, made of platinum of course, which make connections, just like a regular neuron, to other neurons (prosthetic or real). The prosthesis additionally is constructed so as to *change* in the way real neurons do, in response to learning, for example. In short, the prosthetic neuron is functionally equivalent to a real, biological nerve cell.[2]

[2] The prosthetic neuron owes some of its plausibility to the fact that neurons perform part of their signal transmitting functions by processes nicely emulated by electrical circuitry (in fact, circuit diagrams are familiar expository models of ionic transport in neurons). But of course, processes in the brain are more complex than those captured by the basic prosthetic neuron, including global effects mediated by hormonal "sprinkler systems". Global functions, and any other missing functional properties of neurons, can be captured either by prostheses sensitive to inputs other than simple presynaptic input (e.g. by a chemical detector which alters prosthetic response in the way a hormone would alter neuron response), or by a secondary prosthetic system, sensitive to the causes of the additional global input. The technical issues are boggling, but the point here is conceptual: It involves the physical possibility of constructing a mechanism which is functionally eqivalent to an individual nerve cell, but which is intrinsically nonbiological.

With that tiny machine in hand, we turn to a brain. For the sake of argument, let it be the reader's brain. Let us, with all the art of future neurosurgery on call, replace just one of your neurons with a prothetic neuron. We will do so carefully: Wherever your neuron receives inputs from other neurons, so will its replacement. And whatever outputs your old neuron had, the shiny new one will too. On the assumption that the prosthetic neuron is functionally equivalent to a normal neuron, and that we have preserved its interconnections in your brain (and that you are protected from tissue rejection, etc.), it is clear that your brain will continue its normal functioning, even at the intercellular level.

The question is, will you have the same *mind* after the prosthetic neuron implant? Your modified circumstance will be analogous to that of someone with a prosthetic hip joint. If the prosthesis is functionally equivalent to the joint it replaces, then all the functions normally involving the hip joint – walking, running, sitting, etc. – continue in spite of the new plastic and metal involved. With your brain it is no different: Cell by cell, nothing has changed. *Whatever* intricate processes underlie the perception of this sentence (for example), those same processes will go on in your brain with the prosthetic neuron implant. By the same argument, consciousness continues despite the scrap of silicon and platinum just over your pineal gland.

The next step in the thought experiment is simple: Replace another neuron with another prosthesis, again preserving all the functional specification of neuron function. And another, and another . . . With each replacement, the same argument holds: If the brain was conscious in the first instance, the introduction of the prosthesis does not change that fact.[3] Repeat the process a few hundred billion times, and you will find yourself with a *prosthetic brain*, a wonder of inorganic mechanism clunking away at the helm of your biological body. Your gleaming prosthetic brain will be performing exactly the functions of your old gray one. Consciousness, we all assume, is one of those functioning internal processes. It doesn't matter that we can't say which one it is, since the prosthetic brain is duplicating all of the functions of a biological brain. Thus it is conscious – or perhaps we should say, *you* are conscious.

To drive home this point, consider your reply to someone who argued that your prosthetic brain, because it was not a "real" brain, could not house a mind or consciousness. You, of course, know subjectively that there is something that it is like to be you, but your opponent dismisses your claims to that effect as "mere output". You must resort, then, to *public* criteria to convince him. But your case is hopeful, because you have all the public signs of mindedness that anyone else has: exposed to questions, jokes, pinches,

[3] The apparent similarity to the "bald man fallacy" is only superficial. The argument is not that one single neuron is *too small* to make a difference, fallaciously extended to the set of all neurons in a brain. Rather, we can allow the neuron replaced to be crucial to the brain or mind. The point is that the brain-plus-prosthesis preserves *whatever* structure and function the original brain had.

tickles, and the sublime, you respond just like a conscious person. Again, whatever your opponent's standards for ascribing consciousness are, you and your prosthetic brain meet them.

The point of the elaborated example is that it is *conceivable* that a machine (in this case, an assembly of prosthetic neurons) could be conscious. But doubtless prosthetic brains are not *technically* feasible at present. A few technical developments, however, do suggest that prosthetic brains – and artificial minds – may not be that remote.

The first relevant innovation concerns computer hardware, and that is the pursuit of parallel processing. Experimental machines are departing from the serial architecture of the Von Neumann computer, which processes only a very few commands during each machine cycle, in favor of machines which process thousands or even millions of basic commands simultaneously. In moving toward massive parallelism, computers converge on an important physiological aspect of brains, which are massively parallel biological processors. The researchers active in parallel processing are not only aware of this convergence of machine and brain, but deliberately searching for ways to apply results in one study to those of the other (Hinton and Anderson 1981).

The second relevant movement is that of contemporary cognitive psychology. Psychologists and philosophers of mind now widely embrace *computational* models of the mind, wherein minds are understood as embodied in devices (like brains) which perform computations with symbols (in a language of thought or "mentalese"; see Fodor 1975, 1981). The model has been profoundly influential in linguistics (after Chomsky), in vision research (Marr 1982), and in other areas of cognition. Significantly, cognitive psychologists have begun to take their computational models to the study of consciousness, analyzing it in terms of simpler functional processes of the brain, conceived as a special sort of biological computer (Dennett 1978; Marcel 1983; Shallice 1972).

One of the striking aspects of the cognitive approach to psychology follows from the nature of computation itself: computational processes are "instantiation independent", meaning that they can be carried out in a variety of media. For example, despite their differences, both pocket calculators and Ph.D.'s can execute the computationally identical process of addition. One way of looking at the computational approach to mental processes, then, is that it is the approach which consists in specifying the program for performing a mental operation. Program in hand, many different types of systems could perform the operation, some biological and some not. This observation applies equally to the cognitive psychology of consciousness: If cognitive models of consciousness turn out to be true, then it will be a short step to take those human "programs" and translate them into a computer language.

For these reasons, then, I submit that artificial minds may be neither logically nor technologically impossibile. The question is, *should* we seek to construct such machines? In order to answer this, let us examine the several relationships between consciousness and value.

2. *Consciousness is the Foundation of Intrinsic Value*

Each of us is a locus of intrinsic value. But on what does this fundamental principle rest? I will suggest here that it is because we are conscious that each of us has value as a Kantian end-in-itself. Though this is an intuitive meta-ethical principle, it stands in need of clarification.

To assert that consciousness is a ground of intrinsic value is not merely the epistemic assertion that in order to *perceive* value (either in oneself or in others) one must be conscious. Rather, there is a stronger and more basic claim, namely that having the capacity for being conscious entails the possession of intrinsic value. The received standards of clinical death reflect this principle by linking death with the cessation of brain function. A body otherwise intact and functional, but without a functioning brain, is regarded as dead. By the same principle, we seek to preserve lives as long as brain function is indicated, even in the wake of severe disruptions of function of the rest of the body.

A few hypothetical cases reveal the relation of consciousness to the value of life. Consider these two possibilities: In situation A, the unfortunate Smith slips, in 1985, into complete unconsciousness, a dreamless coma, but otherwise remains functioning and "healthy". With intravenous feeding and hospital care he lingers in that quiescent state for three decades, and then dies. In situation B, Smith escapes his coma. He leads a normal life for ten years, until 1995, and then dies. In A, then, Smith lives an extra twenty years. But in B, he enjoys ten extra years of consciousness. Which is the morally preferable situation? Whether we take Smith's point of view, or look upon it from the outside, the same conclusion is evident: we seek to prolongand preserve consciousness over life.[4]

From premises 1 and 2 we arrive at the following conclusion: *Prosthetic brains can have real value*. That is, we can mechanize consciousness without undermining the intrinsic values conscious beings possess. A review of the original prosthetic brain case underlines this: whatever value inheres in you before the wholesale neuron replacement, that value endures thereafter.[5]

[4] In "Death", Thomas Nagel agues that "a man is the subject of good and evil as much because he has hopes which may not be fulfilled, or possibilities which may not be realized, as because of his capacity to suffer and enjoy" (Nagel 1979). His argument turns on cases in which harms occur to individuals without their awareness of them. But consciousness is not merely the passive recipient of experience, but the active initiator of hopes and possibilities. To turn Nagel's cases around, imagine a brilliant accomplishment brought off, somehow, while sleepwalking. The absence of ordinary consciousness undermines the value, and agency, of any action.

[5] The original prosthesis example describes the production of a *copy* of a particular brain, but this feature is inessential. Perhaps in rewiring your brain, several connections were changed, leading to deep alterations in your personality. In spite of this, your fundamental ethical status remains unaltered – this case is comparable to cases of brain injury. And, even if a prosthetic brain were *sui generis*, without model, the fact of consciousness would still establish the fact of intrinsic value. New consciousnesses, minds without precedent, appear every day in maternity wards.

The value inherent in the capacity for consciousness is quite general. It underlies, for example, the prima facie moral wrongness of causing pain. Note also that it is consciousness, not intelligence, that matters most. To lose one's memory or intelligence is a terrible loss, but secondary to the loss of the ground of both intelligence or memory, consciousness.[6]

Now we are in a position to examine a counterargument from metaphysical dualism. The dualist might argue as follows: "The seat of value is indeed the mind and consciousness, but these do not inhere in physical things (neither regular nor prosthetic brains). Rather, mind is a separate substance, and the moral status of a human being rests there – in his or her soul – rather than in any physical trapping." Let us set aside the intractable metaphysical problems with this position and simply assume that it is true. The dualist will want to conclude that normal, biological brains have souls in them while prosthetic brains do not. But he has no ground for making this distinction: the sundering of mind and brain leaves open the Dantesque possibility that living among us are biologically normal people *without* souls, consciousness, or mind (*Inferno* XXXIII, 115 ff.). And it leaves open the possibility that a prosthetic mind, or even a tree or typewriter, could be endowed with a soul or mind. In neither case could we ever tell with certainty what is minded and what is not, since the separate substance leaves no palpable traces of its presence. In our state of ignorance, we must do the best we can, which means giving the benefit of the doubt to *apparently* minded entities in our moral deliberations. We are left, then, in the same position of extending moral stature to prosthetic minds along with ordinary ones.

The disgruntled dualist may attempt to posit mind-body interaction to save him from ignorance, allowing minds to be manifest in, for example, actions. From action, according to the interactionist view, one infers the existence of mind, eliminating bizarre cases like the miraculously minded typewriter. But the prosthetic brain case covers this possibility, since the prosthetic brain issues in the same sorts of actions and utterances as the real one. Whatever the dualist's standards for mindedness, a prosthetically minded person will meet them.

3. *We Are Obligated Toward Artificial Minds.*

So far I have argued that prosthetic brains are conceivable and that they are "artificial minds" possessing whatever consciousness real minds have. And I have suggested that by virture of being conscious, such minds acquire the deeply rooted intrinsic value that human minds have. In what does this value consist? Or, in other words, what rights do artificial minds have? What obliga-

[6] One could argue that the value conferred by consciousness extends not only to artificial minds, but to some higher animals as well – yet we routinely kill animals, suggesting that for some reason *their* consciousness doesn't matter, and thus, perhaps, that value resides not in consciousness but in some other capacity. But killing animals can be justified without abandoning the view that consciousness founds value. Animal killing may reflect a hard choice among values, of the sort discussed following premises 3, 4, and 5 below. Alternatively, it may be that we are often wrong to kill (conscious) animals.

tions do those of us with natural minds possess toward artificial minds?

The straightforward answer is that artificial consciousness possesses the rights of natural consciousness, and our obligations toward it are what they are toward our fellows. This follows directly from premises 1 and 2, and it is hard to see how it could be otherwise: On what grounds could one morally *distinguish* the two sorts of mind? The case stipulates strong equivalences between the functional structure of the two brains. The fact that they are differently *embodied* is properly irrelevant: if a human being were *born* with a radically different structure from the ordinary, we would not reject his or her claim to rights. We correctly seek to be morally blind toward many of the details of an organism's makeup. If I am right, then what we principally attend to is that quality which appears equally in real and prosthetic brains, consciousness.

The specific rights and obligations borne by the minded are not at issue here. They are probably close to the standard canon: Minds, artificial or otherwise, are entitled (prima facie) to continued consciousness, and entitled (prima facie) to the furtherance of whatever they undertake, provided their projects do not conflict with similar rights of others. Artificial minds give rise to a few special claims. Here the crucial fact is that these minds are *created*, by which fact the creator acquires special responsibilities. Like a parent, the creator is specifically obligated to preserve and enrich the life of his or her creature. That means, among other things, that the created being can claim his, her, or its rights specifically from the creator. Again, the point is not to take a stand on issues of right and obligation. Rather, it is to urge that whatever one's morality entails concerning conscious *people*, it will similarly entail toward conscious *machines* (On obligations of parents, see O'Neill and Ruddick 1979, section II).

4. *Differently Embodied Minds Have Different Subjective Values*

This premise takes the argument in a new direction, turning from the issue of intrinsic value to that of assumed or subjective values, i.e. the values and goals chosen and pursued by different sorts of minds. We move, in short, from value treated in general and from the outside, to its specifics, viewed from within. The fact of obligation makes this an important issue: we are obliged, following the discussion above, to further the projects of artificial minds just as we are obliged toward regular minds.

But what projects are these? We humans share general aims, a small set of basic goals which are common to all of us – food, clothing, and shelter are among the immediate examples. Would an artificial mind also seek these goods? We just don't know. This in itself raises the serious possibility of differing basic aims. Although intrinsic value travels with consciousness independently of the details of the instantiation or means of embodiment of consciousness, subjective value varies widely according to the constitution of organisms.

For example, let us return to the case of a prosthetic brain. We have care-

fully captured the functional capacities of the biological brain, and so one would expect that no changes would occur in the values of the person whose brain was so replaced. But suppose the prosthetic brain, unlike a biological brain, will never wear out. Without altering any other detail of the case, the subject's choices could be reasonably expected to shift dramatically. For example, she no longer needs to worry about mental decline with age. This offers the possibility of many more productive years, which increases the time available for major projects. At the same time, her attitude toward the rest of her body alters: it is now the sole obstacle between her and immortality, a too solid weight on her mind. If her brain's functioning depends on the body, then she will probably seek to protect and preserve her body with zeal, or to have it replaced with prosthetic parts.

When we turn to artificial minds with radically different embodiments, the differences from the subjective values embraced by humans increase. What would be the goals of a mind equipped with a body impervious to the elements, or resistent to specific toxins or radioactivity, or responsive to X-rays but not to light, or accessible only through keyboards? What would the standard of beauty be to minds typically housed in gray boxes? What would communication or honesty mean to creatures who can transmit the entire contents of their minds to their fellows in seconds? What would community mean when identical minds can be fabricated magnetically, with no need for nurturing or rearing? What would morality include for minds immune from pain or natural death? What can we expect from minds who exceed our intelligence, by our own standards, by an enormous amount? We can speculate about all of these, but at bottom we can expect a deep and imponderable gulf between us and them.

In stressing the different values of differently embodied consciousness I depart from an old assumption of Promethean AI, namely, that the created mind seeks whatever its programmers program it to seek. It is true that every computer program to date works toward narrow "goals" prescribed by human fabricators. But if a future AI machine really did capture the human abilities of self-reflection and analysis, we can expect it to be like us: variable in its goals and allegiances, subject to the thousand natural and artificial shocks that flesh and silicon are heir to.

Even if reliable control of subjective value were possibile, one must ask whether such control is ethical. Just turn the tables: Would it be ethical for an outside agent to *install* the deepest goals and values in *us*? Such a deed becomes monstrous as those installed aims diverge from the aims which are natural for each of us to hold.

5. *In a Collision of Values, Human Values Must Be Chosen.*

Many different agents in a community of minds – whether natural or artificial – will hold many differing subjective values. This pluralism can be a good thing, adding to the depth and richness of a society. But in some cases, values may collide, particularly judgments concerning the course of a com-

munity or society as a whole. If such a collision occurs between artificial minds and human minds, whose values are to be chosen? I believe that the ethical choice will always be the human one. While I think a positive case can be made for this claim, in fact I will argue something different, namely, that to chose the values of the nonhuman group is, in a sense, incoherent.

The argument for the incoherence of chosing nonhuman values is simple. First, consider some of the entailments of embracing a value: for such a choice to be rational, the chosen value must be consistent by and large with the other held values. And for the whole set of held values to be rational, they must be appropriate to the beings who hold them. "Appropriate" here is unavoidably vague, but some of its sense was communicated in the discussion of the fourth premise. Appropriate human values, for example, are those which can be realized by human minds and human bodies. They are the values which promote their own preservation, and promote the recognition of the care-worthiness of other humans. For example, an ethic of treachery is not, in the end, a rational value choice, as Kant and his followers detail. Similarly, to embrace the aim of breathing underwater, or perpetually standing on one's head, is also inappropriate for simple physiological reasons.

With that understanding of rational value choice, we examine the possibility of chosing a nonhuman value. If such a choice is rational, then it meets the (admittedly rough) standard of appropriateness *for humans*. But if that is so, then nonhuman and human values can coincide, and the nonhuman value is no longer strictly speaking a nonhuman value. The differences have been ironed out and the collision avoided. On the other hand, if the candidate value truly is nonhuman, that is, appropriate to beings of fundamentally different natures, then we cannot rationally embrace it – it cannot be appropriate for us.

Thomas Nagel makes vivid the paradox of embracing fundamentally outside values is his discussion of the meaning of life in "The Absurd":

> A role in some larger enterprise cannot confer significance unless that enterprise is itself significant. And its significance must come back to what we can understand, or it will not even appear to give us what we are seeking. If we learned that we were being raised to provide food for other creatures fond of human flesh, who planned to turn us into cutlets before we got too stringy – even if we learned that the human race had been developed by animal breeders precisely for this purpose – that would still not give our lives meaning, for two reasons. First, we would still be in the dark as to the significance of the lives of those other beings; second, although we might acknowledge that this culinary role would make our lives meaningful to them, it is not clear how it would make them meaningful to us. (Nagel 1979, p. 16)

Just as it cannot make sense to the chicken to embrace the chicken-destroying projects of human beings, so it cannot make sense for humans to embrace the inscrutable projects of nonhuman minds.

Of course, that goes the other way as well: If artificial minds are conscious,

intelligent, and rational, we can expect them to draw the same conclusions about their values as we do about ours.

6. *On Balance, We Should Not Pursue Artificial Minds.*

There are several uncertain points in the five premises I've discussed so far, but on balance I think they recommend, on ethical grounds, that we be very cautious about the pursuit of cognitive simulation AI. The main reason is that the pursuit of conscious machines seems to lead to situations where ethical violations necessarily occur.

The violations emerge as the potential consequences of premises 3, 4, and 5. In a nutshell, if we pursue artificial minds, we incur obligations to foster their particular values. But those values are likely to diverge from our own, in profound and presently unfathomable ways. In the face of a collision of values, we must rationally and naturally chose our own values, thus violating the prima facie rights of the conflicting artificial minds.

There is a utopian vision of the future of AI, including the creation of artificial minds, in which the collision of values does not occur: The artificially intelligent machines imagined in that vision are friendly, kind, obedient, and wise; their provident management of complex human affairs creates a global playground for their (characteristically short-sighted, error prone, and aggressive) human creators.[7] But those utopian outcomes are unlikely. Instead, we are likely to preside over the creation of fundamentally alien beings, whom we will unwittingly manipulate and abuse. With that possibility in mind, the creation of consciousness looks like an ethically dubious enterprise.

The price of turning back from cognitive simulation AI is a loss of potential knowledge, and the area where this loss would be painful is in the knowledge of human consciousness. If the discussion of premise 4 is correct, however, creating conscious machines will not teach us much about human consciousness – we will always wonder whether what we observe is at all like what happens in people. We are left with basing the study of human consciousness on human beings, and this seems like a reasonable and fruitful enterprise.

We will miss, then, the opportunity to observe alien minds in action. Thus, some knowledge will be lost. But the lesson of Frankenstein – and the mushroom cloud – is that some knowledge comes at a price too high.[8]

University of California
Santa Barbara, CA 93106 USA

[7] See McCorduck 1979, especially Chapter 14. Even if the utopian possibilities could be realized, would we *want* such a world? Isn't the management of complex human affairs really part of *human* responsibility? The issue of responsibility and its abdication is crucial to the future of AI, and many authors discuss it. See especially Weizenbaum 1976.

[8] Work on this article was supported by a University of California Regents' Junior Faculty Fellowship. Thanks are also due to Terrell Ward Bynum for his helpful comments.

References

Dennett, D. C. (1978) "Toward a Cognitive Model of Consciousness", In *Brainstorms*. Montgomery, VT: Bradford Books.
Fodor, J. A. (1968) *Psychological Explanation*. New York: Random House.
Fodor, J. A. (1975) *The Language of Thought*. New York: Crowell.
Fodor, J. A. (1981) *RePresentations*. Montgomery, VT: Bradford Books.
Hinton, Geoffrey E. and Anderson, James A., eds. (1981) *Parallel Models of Associative Memory*. Hillsdale, NJ: Lawrence Erlbaum.
McCorduck, Pamela (1979) *Machines Who Think*. San Francisco: Freeman.
Marcel, Anthony J. (1983) "Conscious and Unconscious Perception: An Approach to the Relations Between Phenomenal Experience and Perceptual Processes." *Cognitive Psychology* 15: 238–300.
Marr, David, (1982) *Vision*. San Francisco: Freeman.
Nagel, Thomas (1979) *Mortal Questions*. Cambridge: Cambridge University Press.
Newell, Allen (1973) "Artificial Intelligence and the Concept of Mind." In Schank and Colby 1973.
Nilsson, N. J. (1971) *Problem Solving Methods in Artificial Intelligence*. New York: McGraw-Hill.
O'Neill, Onora and Ruddick, William, eds. (1979) *Having Children*. Oxford: Oxford University Press.
Schank, R. C., and Colby, K. M., eds. (1973) *Computer Models of Thought and Language*. San Francisco: Freeman.
Shallice, T. (1972) "Dual Functions of Consciousness," *Psychological Review* 15: 383–393.
Turing, A. M. (1950) "Computing Machinary and Intelligence." *Mind* 59.
Weizenbaum, Joseph (1976) *Computer Power and Human Reason*. San Francisco: Freeman.