Artificial Intelligence and Ethics: An Exercise in the Moral Imagination

Michael R. LaChat

The Methodist Theological School in Ohio, Delaware, Ohio 43015

In a book written in 1964, *God and Golem, Inc.,* Norbert Wiener predicted that the quest to construct computer-modeled artificial intelligence (AI) would come to impinge directly upon some of our most widely and deeply held religious and ethical values. It is certainly true that the idea of *mind as artifact,* the idea of a humanly constructed artificial intelligence, forces us to confront our image of ourselves. In the theistic tradition of Judeo-Christian culture, a tradition that is, to a large extent, our "fate," we were created in the *imago Dei,* in the image of God, and our tradition has, for the most part, showed that our greatest sin is pride—disobedience to our creator, a disobedience that most often takes the form of trying to be God. Now, if human beings are able to construct an artificial, *personal* intelligence—and I will suggest that this is theoretically possible, albeit perhaps practically improbable—then the tendency of our religious and moral tradition would be toward the condemnation of the undertaking: We will have stepped into the shoes of the creator, and, in so doing, we will have overstepped our own boundaries.

Such is the scenario envisaged by some of the classic science fiction of the past, Shelley's *Frankenstein, or the Modern Prometheus* and the Capek brothers' *R.U.R.* (for Rossom's Universal Robots) being notable examples. Both seminal works share the view that Pamela McCorduck (1979) in her work *Machines Who Think* calls the "Hebraic" attitude toward the AI enterprise. In contrast to what she calls the "Hellenic" fascination with, and openness toward, AI, the Hebraic attitude has been one of fear and warning: "You shall not make for yourself a graven image..."

I don't think that the basic outline of *Frankenstein* needs to be recapitulated here, even if, as is usually the case, the reader has seen only the poor image of the book in movie form. Dr. Frankenstein's tragedy—his ambition for scientific discoveries and benefits, coupled with the misery he brought upon himself, his creation and others—remains the primal expression of the "mad scientist's" valuational downfall, the weighting of experimental knowledge over the possibility of doing harm to self, subject, and society. Another important Hebraic image is that of *R.U.R.,* a 1923 play that gave us the first disastrous revolt of man-made slaves against their human masters. In both works theological insight and allusion abound; God and creation are salient issues, and both works condemn the AI enterprise.

Alongside the above images, of course, there have always lurked Hellenic or, perhaps better, "Promethean" images. Western history is replete with examples of attempts to construct an AI, with miserable and comical flops; with frauds; and with, as of late, some feeble approximations. The more sophisticated our reason and our tools, the more we seem to be inexorably drawn to replicate what has seemed to many to be what marks human persons off from the rest of creation—their *cogito,* their *nous,* their reason. We seem to want to catch up with our mythology and become the gods that we have created. In the layperson's mind, however, the dominant imagery appears to be the Hebraic; many look upon the outbreak of AI research with an uneasy amusement, an amusement masking, I believe, a considerable disquiet. Perhaps it is

Abstract

The possibility of constructing a *personal* AI raises many ethical and religious questions that have been dealt with seriously only by imaginative works of fiction; they have largely been ignored by technical experts and by philosophical and theological ethicists Arguing that a personal AI is possible in principle, and that its accomplishment could be adjudicated by the Turing Test, the article suggests some of the moral issues involved in AI experimentation by comparing them to issues in medical experimentation Finally, the article asks questions about the capacities and possibilities of such an artifact for making moral decisions It is suggested that much *a priori* ethical thinking is necessary and that that such a project cannot only stimulate our moral imaginations, but can also tell us much about our moral thinking and pedagogy, whether or not it is ever accomplished in fact.

the fear that we might succeed, perhaps it is the fear that we might create a Frankenstein, or perhaps it is the fear that we might become eclipsed, in a strange oedipal drama, by our own creation. If AI is a real possibility, then so is *Frankenstein.* McCorduck says of *Frankenstein* that it "combines nearly all the psychological, moral, and social elements of the history of artificial intelligence."

Philosophical and theological ethicists have been silent, with a few exceptions (Fletcher, 1977), on the problem of AI, leaving, unfortunately, those with little training in ethical theory to assess the moral arguments. Boden (1977), Weizenbaum (1976), McCorduck (1979), and Hofstadter (1980), among others, have dealt with questions of technique, with the "hardware and software" questions surrounding the possibility of AI. Even when such researchers and chroniclers consider ethical questions, they tend to focus on the effects of AI upon society and not upon the AI *qua* subject of experimentation. By focusing on the morality of such experimentation as it effects the subject, I am obviously entering into the realm of the moral imagination, a realm that most ethicists might find trivial, farfetched, or meaningless given the present problems of the planet. Imagination *per se* has often been neglected in philosophy and theology, and the moral imagination suffers more neglect than the ordinary kind, perhaps because it seems more playful than the austere and often overly sober undertakings of most ethicists. Having team taught, however, a course on artificial intelligence and morality, a course about which I was, at first, somewhat dubious, I have reached the conclusion that pedagogically it is a very productive issue for ethical thinking, whether it will ever be accomplished in fact. The problems involved in the construction of a person by artificial means are fascinating and provocative partly because they allow us distance on ourselves; they allow us to probe ourselves in ways our imaginations were previously limited in doing This does not mean, however, that they do not pose serious problems for the moral imagination.

One does not have to be a theist in order to be able to distill some practical wisdom from the religious counsel, "Don't play God " Among other things, God is a moral concept, as Kant rightly asserted. This venerable injunction can be "demythologized" to a word of warning, of caution toward all human undertakings, the effects of which might be irreversible or potentially harmful to ourselves and to others. For some ethicists, myself included, the first word of ethics is identical with the first caution of ethical medicine—"Above all, do no harm." As W. D. Ross (1965) has pointed out, such negative injunctions are the guiding thoughts, indeed, are the form of almost all legal codes, primitive or modern The *prima facie* duty of nonmaleficence, of not doing harm, is almost universally conceded to be more stringent than positive injunctions to "do good." In the language game of modern ethics, this latter injunction can be considered as part of the utilitarian tendency to explore the new, to take risks that might even cause harm to a few for the benefit of many.[1] It is certain that both injunctions can claim to be moral, but I side with Kant on the primacy of the former: All rational beings, capable of moral evaluation, must be considered as ends in themselves rather than as means to another's ends.[2] The stringency of the norm of nonmaleficence, attested to in almost all of the modern codes of experimental medicine, means that ethical thinking with regard to the possibility of constructing an artifact which might verge on the personal is necessary *a priori.* The intent of this article is, thus, to raise the questions of a moral nature by stimulating the imagination in tandem with the technological imagination, a necessary reciprocal relationship that we cannot allow to be submerged entirely in the technological realm.

In the first part of the article, I argue briefly that replication of personal intelligence is possible in principle, because counterarguments usually rest on some sort of quasi-mystical dualism, which I find untenable. I further argue that the Turing Test allows us to specify the conditions under which a machine could be said to have attained "personhood," however difficult such a characteristic might be to define.

In the second part of the article, I ask whether such an undertaking should be pursued. The focus of this section is on the moral safeguards for the subject of the experiment, and questions are raised about the extent to which an analogy can be drawn between the morality of the AI project and the ethical guarantees given to human subjects by modern experimental medicine.

The last section of the article is a true exercise in the moral imagination. It asks the question, "Can an artificial intelligence be moral?" It is suggested that one cannot answer this question without giving consideration to the perennial philosophical problems of free will, casuistry, and the role of emotions in moral decision making.

Is Artificial Intelligence Possible in Principle?

I will not pretend in what follows to be an expert on the "hardware" or "software" of AI, nor do I claim to be particularly adept in the relevant areas of philosophy of the mind. These technical questions have been dealt with elsewhere in great detail.[3] Joseph Weizenbaum, an increasingly isolated figure within the relatively small circle of AI researchers, has rightly pointed out some of the enormously exaggerated claims that even a sophisticated audience is prone to make for a technology it really does not understand. Few people who have been exposed to the relevant literature would doubt the incredible complexity of

[1]For a more thorough discussion of normative theories, see W K Frankena's (1973) *Ethics,* chapters 2 and 3

[2]See Paton (1965), especially chapter 16

[3]For example, McCorduck (1979), pp. 359-64

such an undertaking. Thus, for example, we need to cautiously remind ourselves that there is a distinction between intelligent behavior and personally intelligent behavior. A machine can learn to run a maze as well as a rat and at the level of rat intelligence could be said to be fairly "smart." The intelligence of the human brain, however, is of a different magnitude entirely.[4] Although it is obvious that machines can perform some activities at a higher level than persons can, these tasks remain, by and large, highly specialized and therefore remote from the capacity of human intelligence for multipurpose activities.

The Obstacle of Dualism

In spite of these difficulties—organizational complexities that might prove to be insuperable—I feel it necessary to stick out my layperson's neck and offer a tentative argument that an artificial, personal intelligence is possible in principle. Having team taught a course titled "Minds, Machines, and Morals," with a mathematician at St. John's University in Minnesota, I am well aware of the skeptical wall that confronts enthusiastic AI researchers. The first response of most students was a flat rejection of the possibility of such an endeavor. As the course progressed, however, such absolute skepticism gave way to a more tempered doubt, namely, that AI would never be accomplished, given the complexity of the task. What occurred to us during the process of in-class debate was that the absolute skepticism rested ultimately on some sort of dualism between mind and brain; for example, it was often contended that persons had "souls" which were not dependent upon the brain for their existence or that persons were possessed of some almost magical "substance" which could not be duplicated artificially.

I happen to agree with Carl Sagan who stated on a public television show that the evidence for such dualism is nonexistent. Although we might not, as the psysiologist Wilder Penfield (1983) suggests, have all of the evidence which would allow us to deny that a "creative thought" might precede electrical activity in the brain, although we may not be able to accept without reservation the philosophic claim that thought equals brain activity, we can make, with some assurance, the more modest claim that there is no evidence of thought taking place without the brain. We do not have to make the stronger claim that there is an identity, an ontological identity, between consciousness and an electrical thought pattern in the brain in order to reject the claim which asserts that conscious thought can occur without the brain.[5]

Let me propose a rather simple response to the absolute skeptics who rest their arguments on an indemonstrable, ontological dualism between mind and brain. All that is necessary to indicate the possibility of AI is to posit a functionally isomorphic relationship between a neural network in the human brain and (at present) a silicon chip in a computer What is asserted here is that intelligent thought is dependent for its existence on the neural "machinery" of the brain, on the flow of electricity through that "hardware." Electrical patterns in the brain can be compared to the software of the computer and the brain's neurons to the computer's hardware, to the "neural" networks of the chip This is not to say that mind cannot be an emergent property of a certain level of organization but only that such emergence would be dependent on a neurological "substrate" for its existence. As Douglas Hofstadter says in his magnificent *Gödel, Escher, Bach*:

> Crucial to the endeavor of Artificial Intelligence research is the notion that the symbolic levels of the mind can be "skimmed off" their neural substrate and implemented in other media, such as the electronic substrate of computers To what depth the copying brain must go is at present completely unclear. (P. 573)

This last sentence is an important one. There are two basic approaches to it in the history of AI research. The first is the cybernetic model; it relies on the "physiological" similarity between neurons and hardware. Thus, Norbert Wiener (1964) can say that even the living tissue of the brain is theoretically duplicable at the molecular level by suitable hardware materials. The second approach, the one dominant today, is the "information-processing" model, a model that does not pay as much attention to the level of hardware similarity. All it asserts is that a machine will demonstrate intelligent behavior when it acts intelligently by behavioral definitions. An intelligent machine would not have to "look like" an intelligent human; it would only have to exhibit the same sort of behavior. It is evident that the functionally isomorphic relationship between neural network and hardware is an admixture of both models, although it leans in the direction of the latter. It claims that the neural network hardware isomorphism is a cybernetic one insofar as the substrate conditions would probably have to be similar enough to facilitate the duplication of the functions of both; it does not claim that the isomorphism would necessarily be pictorial as well. What we would need, then, in order to duplicate the hardware substrate would be an adequate "map" of the brain and the extremely sophisticated "tools" and "materials" with which to duplicate it.

---

[4]It had been estimated that the human brain contains 10^{13} bits of intelligence capacity as opposed, for example, to a cow's capacity of 10^{11} bits (*Chemtech*, 1980, p 590)

[5]As Karl Marx and Friedrich Engels pointed out, mind-body dualists tend to hold to a disconcerting dualism between their physics and metaphysics as well Thus, they say rightly of Descartes: "Descartes in his physics endowed matter with self-creative power and conceived mechanical motion as the act of its life He completely separated his physics from his metaphysics Within his physics matter is the only substance, the only basis of being and of knowledge" (Marx & Engels, 1971, pp 60-61)

So far I have only discussed conditions relevant to the hardware substrate. The software level of a personal machine consciousness—the symbolic level that Hofstadter contends must be "skimmed off" the neural substrate and implemented in other media—seems to be much more problematic to duplicate. A personal intelligence must have personality, and this seems on the face of it to be an almost impossible problem for AI. Personalities are formed through time and through lived experience, and the personal *qua* humanly personal must certainly include the emotional. Indeed, a phenomenological analysis of human experience, such as that of the early Heidegger, indicates that persons might have to experience the emotion of dread in the face of finitude (death) in order to have a grasp of "isness," in order to be fully conscious (Heidegger, 1962; Dreyfus, 1972).

The temporal dimension of human consciousness is a great obstacle to the AI project. Suppose we were able to produce a "map" of the thought patterns of a human adult. Such a map would have to include memories and experiences as well as hopes, aspirations, and goals; in short, it would have to be a map inclusive of the three temporal dimensions of human consciousness—past, present, and future. Could such a map be "programmed" directly into the hardware substrate? The question of the growth of consciousness through time thus emerges as a particularly salient problem. Perhaps a personally intelligent machine has to grow into consciousness, much as a human baby does; then again, perhaps not.

It is not out of the realm of possibility that pain consciousness might be electrochemically duplicable. To what extent are pain and emotion necessary to personal intelligence? There are certainly cases in which paralyzed or drugged persons experience no pain and yet are still conscious, but such persons might still be said to suffer. This crucial distinction (Boeyink, 1974) needs to be borne in mind. Suffering, as opposed to pure pain reflex, is strained through a human ego, an ego that can remember, anticipate, and project itself in time. Perhaps, then, emotions and body consciousness are indispensable requisites of the personally intelligent. These are difficult issues that probably can only be resolved through the trial and error of experiment—and there's the rub! The deliberate constructions of the capacity for pain would seem to entail the causation of pain as well (as the robots in *R.U.R.* are given pain in order to keep them from injuring themselves). This problem raises the question of the morality of the experiment in a vivid way, and is an issue I address shortly.

Adjudicating the Achievement of a Personal AI

If these difficulties can be surmounted, how can we finally reach agreement that a personal AI has been achieved? The quasi-behavioral criteria I outlined earlier facilitate our ability to specify the conditions under which a machine can be said to have achieved this intelligence, namely, the conditions specified by the Turing Test.

The simplest way to adjudicate a claim that a machine has or has not achieved personal intelligence is to define intelligence behaviorally and then test the machine to see if it exhibits it. No doubt, such definitional lists, even when specified exhaustively, would still run afoul of persistent human skepticism, but such skepticism can be greatly alleviated by the application of the Turing Test. Alan Turing, a British mathematician and logician who believed in the possibility of AI, proposed a test that might give criteria for ascertaining the accomplishment of such an AI (McCorduck, 1979). Suppose that a person (an interrogator) were to communicate unseen with both another person and the AI "subject," and that the interrogator could not tell whether he or she was communicating with the other person or with the "machine"? I contend that such a test could adjudicate the claim, provided (1) that we could agree at the onset with what the test presupposes about the normatively personal, and (2) that certain problems concerning the prerequisites of the interrogator could be ironed out.

It is evident that the test presupposes communication as the *sine qua non* of personal intelligence, that the ability to meaningfully converse through the medium of some sort of language constitutes the essentially personal. Such a presupposed concept of the personal is evident in much of the literature of current medical ethics. Because communication depends on the functional organization of the brain, brain death might be considered the termination of the personal (Fletcher, 1977). Some philosophers and theologians have argued cogently that an inability to respond to stimuli, particularly to linguistic communication, means the subject, although perhaps morphologically human, is no longer a person.[6] Theologians, in particular, are apt to define the personal with regard to the giving and receiving of "the Word" (Niebuhr, 1963). Whether we take such communication to be exhaustive of the personal, it is apparent that without its possibility persons would no longer be persons. The high level of hardware and software sophistication necessary to enable symbolic communication ought to encompass any other kinds of activity we might take to be personal.[7] Human beings do not become persons through simple biological conception. A zygote is a human being, but it can only trivially be considered a person. A machine can be a person in the same way that a "higher" animal might possibly be considered a person (Singer, 1980)—if it shows the ability to

[6]For example, Joseph Fletcher (1972) uses the term "human" rather than "person," but his criteria distinguish the personal from the merely biomorphologically human.

[7]It ought to encompass even the possibility of a "religious experience," which Weizenbaum has asserted in a popular source it could never attain to Weizenbaum does not tell us why it could not have such an experience. See the interview of Weizenbaum by Rosenthal (1983), pp. 94-97

meaningfully participate in a language system.

Other questions about the Turing Test remain. These problems have largely to do with the capacities of the interrogator. Such a person, it appears, must be a rational adult, capable of giving reasons for a decision. Should the interrogator know something about computers as well? Would the interrogator be able to "trick" the computer in ways a layperson could not? Would one of the tricks be to shame or insult the subject in order to elicit an emotional response? Perhaps the best interrogator would be another computer, one with the capacity, for example, to interrogate the computer well enough to see at what point it might eclipse human ability. Should there be a group of interrogators, both human and not? How long should the test run before a decision is made? The computer might have to be capable of deceiving the interrogator and of fabricating a life history (in the supposed absence of having "lived" one). It might need the capacity to anticipate tricks and perhaps even to evince a capacity for self-doubt. In short, it might have to possess a self-reflexive consciousness (the ability to make itself the object of its own thought), a characteristic that Tooley (1974) and Mead (1972) have convincingly argued to be a hallmark of the personal self. Such a machine might even have to be able to lie. An intriguing theory is that a child comes to know himself or herself as an ego when he or she can deceive an adult to whom he or she has previously attributed omniscience; then the child finally knows he or she has a private consciousness (Schlein, 1961). Such might also be the case with intelligent machines.

These problems are indeed difficult, though fascinating. At any rate, having asserted that a machine might in principle achieve a personal consciousness, I find I am still begging the central question of the article. The Turing Test, if passed, would be a *fait accompli*. The harder question is to ask whether it should have been undertaken in the first place.

Is the Construction of a Personal AI an Immoral Experiment?

Listen, for a moment, to the lament of Frankenstein's monster:

> Like Adam, I was apparently united by no link to any another being in existence, but his state was far different from mine in every other respect. He had come forth from the hands of God a perfect creature, happy and prosperous, guarded by the especial care of his creator, he was allowed to convene with, and acquire knowledge from, beings of a superior nature, but I was wretched, helpless, and alone. Many times I considered Satan was the fitter emblem of my condition. For often, like him, when I saw the bliss of my protectors, the bitter gall of envy rose up within me... Hateful day when I received life! ... Accursed Creator! Why did you form a monster so hideous that even *you* turned from me in disgust? (Pp. 152-53.)

It seems obvious, first of all, that a creature who can communicate as well as the monster should have little trouble passing the Turing Test. Second, the fact that the monster is not a "machine" but an assemblage of biological parts revivified by electricity is beside the point. What is intriguing about the monster's lament is that he is claiming something analogous to a "wrongful birth" suit; as an imperfect creation, he is claiming that he ought not to have been made. However fanciful the monster might be, he is a perfect example of what might go wrong with AI: He is the possibility of an experiment gone awry. His story would be incomplete without the rationalizations of his creator:

> I believed myself destined for some great enterprise
> ....I deemed it criminal to throw away in useless grief those talents that might be useful to my fellow creatures.. .all my speculations and hope are as nothing and, like the archangel who aspired to omnipotence, I am chained to an eternal hell. (P. 256.)

We can say the monster is claiming that he was the improper subject of a poorly-designed, nontherapeutic experiment, whereas Dr. Frankenstein claims as his motivation not only his own egotism but the benefits to society that might accrue from his experiment. In modern times, a similar form of utilitarian justification is echoed by Norbert Weiner (1964), who says of the fears surrounding the AI enterprise:

> If we adhere to all these taboos, we may acquire a great reputation as conservative and sound thinkers, but we shall contribute very little to the further advance of knowledge. It is the part of the scientist—of the intelligent man of letters and of the honest clergyman as well—to entertain heretical and forbidden opinions experimentally, even if he is finally to reject them. (P. 5)

Entertaining opinions is one thing, but here we are talking about an experiment, an experiment in the construction of something which is so intelligent that it might come to be considered a person. It is one thing to experiment on a piece of machinery, such as a car, and quite another, as the Nuremberg Medical Trials indicated, to experiment with, or in this case toward, persons . Kant's imperative to treat persons as ends in themselves rather than as means to an end is at the heart of the matter.

Human Experimentation and the AI Experiment

Particularly since the Nazi atrocities, the norm of nonmaleficence has been considered more stringent than that of beneficence in subsequent medical codes. The concern has been overwhelmingly in favor of the subject and

against the interests and benefits of the society. Even where considerations of social utility have been included in such codes, a strong burden has been placed on the experimenter to prove that benefits overwhelmingly outweigh risks to the health and welfare of the subject. In the extreme, such concerns have tended to rule out all forms of nontherapeutic experimentation (Jonas, 1977). Is the AI experiment then immoral in its inception, assuming, that is, that the end (telos) of the experiment is the production of a person?

Let us first ask whether the experiment can be considered therapeutic in any meaningful sense of the word; that is, is it of benefit to the subject? It is difficult to consider it as such, because the subject does not really exist as confirmed fact (as a person) prior to the experiment itself; it exists only *in potentia.* In the stages prior to the "dawning" of consciousness, the machine is in some respects more like a zygote or early fetus than a person (assuming, of course, that the early stages of the construction of the AI are somewhat analogous to the early stages in the teleological process of biological growth). We can, then, consider the experiment to be therapeutic only if we maintain that the potential "gift" of conscious life outweighs no life at all. We cannot say that this is an experiment on a sick person or a sick potential person for the sake of that person whom we are attempting to make better. Thus, we can fairly consider the experiment to be nontherapeutic in nature. If so, the stringent code of medical ethics seems to apply (Ramsey, 1975). We shouldn't treat the subject as if it had no rights at all. This is not to say that the benefits to society would be trivial; even pure knowledge is never trivial, especially when such "high tech" almost invariably trickles down in various ways. However, the presumption of the experimental tradition is always *prima facie* against nontherapeutic experimentation, save perhaps in those cases where the experimenter is also the subject (Jonas, 1977).

It might be contended, however, that the "birthing" of an AI is analogous to a human birthing. Let us consider this possibility for a moment.

Until the recent present in human history, birthing has basically been "set" for us. We took what we got. In the Judeo-Christian tradition, life itself has been viewed, for the most part, as a gift from God, and the sanctity of human life has been weighted very strongly against any quality of life ethic (Dyck, 1977; Noonan, 1970). Modern technology, specifically genetic screening and amniocentesis, has, when coupled with the Supreme Court's abortion decision, raised a different sort of moral question: Is it immoral to knowingly bring into the world a potentially or actually defective child? Such a question raises the crucial issue of the locus of rights in the procreative decision.

Roe vs. Wade has shown that the right to procreate and to terminate procreation (whether or not the fetus is defective) is, at least in the first trimester of pregnancy, a subjective, discriminatory right of the parents (Reiser, Dick, & Curran, 1977). The desires of the parents are the locus of the rights, and only at stages approaching viability can the state take any compelling interest in the welfare of the fetus. Yet questions can be raised about the rights of the fetus. I have in mind here not the right to life of the fetus but the right of the potential newborn to be born to the possibility of a healthy existence, the right of a child to be born without disabling handicaps and with the freedom from severe pain (as opposed to the capacity to feel pain).

It is my opinion that most AI researchers would incline to follow something analogous to the "subjective desires for the parents" route where the possibility of AI is concerned, thereby implicitly making an analogy with procreative rights as guaranteed by the Supreme Court decision. However, the analogy does not really hold for a number of significant reasons.

In a certain sense, we must acknowledge that human reproduction is an experiment. Joseph Fletcher (1972) and others have argued that the "sexual roulette" mode of human reproductive experimentation should come to an end. "To be human is to be in control," says Fletcher, and he argues that a baby conceived by artificial means would be more, rather than less human because it would be the product of a thoughtful, rational process. Nonetheless, we can say that the human "roulette" mode of reproduction has allowed us inductively to generalize certain rules of thumb which serve as guidelines for preventive, prenatal care of the fetus. In AI there are really few, if any, such precedents. The injunction "do no harm" is cloudy in the case of AI, much more so than the injunction, for example, not to smoke while pregnant. AI is an extreme experiment; we have little or no knowledge of what might happen. Although it could be argued that all preventive, human reproductive behavior has been built on the basis of trial and error as well, we can at least say that evolution has set our birth for us in a way which, at present, is only trivially a matter of the will. AI is really much more of a deliberate experiment.

Second and most important, human reproduction is a necessity for species survival. All of the earth's cultures attest to this necessity and provide ethical and legal safeguards to protect it. AI, on the other hand, is not necessary as a means of species survival. This does not mean that it might not prove in ages to come to be a necessity, but at present it is not. AI is more a luxury than a necessity; as such, it should fall under the stringent experimental guidelines. It can also be argued that at present the risks of the AI experiment are greater than the benefits, and the ratio of risk to benefit is higher in AI than in human reproduction. For the *subject* of the AI experiment, the only benefit appears to be, at best, the gift of conscious life. This gift must be weighed against whatever problems might arise.

The result of this argument is, apparently, to side

with the Frankenstein monster's "wrongful birth" suit. An AI experiment that aims at producing a self-reflexively conscious and communicative "person" is *prima facie* immoral. There are no compelling reasons which lead us to believe that we could ensure even the slightest favorable risk-benefit ratio. Is the necessary conclusion, then, not to do it?

Does a Personal AI Have Rights?

Suppose, for a moment, that we could guarantee all of the conditions requisite to birthing an AI which had the full range of personal capacities and potentials; in other words, suppose we could guarantee that the AI would have the same rights *a priori* which actual persons do now. What would such rights be? A suitable starting point might be the United Nation's 1948 Declaration of Human Rights (Donaldson & Werhane, 1979). I excerpt some of those which might be pertinent to the AI case. Article Four, for example, states that no one shall be held in slavery or servitude. Isn't this the very purpose of robotics? Fletcher has already contended that the bioengineering of such entities might conceivably be warranted for the performance of dangerous tasks (an interesting parallel to the conditions making for *R.U.R.*'s robot revolt) (Fletcher, 1972).

The prohibition of slavery raises a further question for AI. A free, multipurpose robot might be "legitimate," as it were, but what of a single-purpose robot? Would this be tantamount to engineering a human baby with, for example, no arms so that he or she could become a great soccer player? Any limited-purpose or limited-capacity AI would have its essence defined before its existence (Sartre, 1957), so to speak; if we would not accept this of the humanly personal, we should not accept it of the AI personal as well. Thus, if we were to hold strictly to the United Nation articles, we would have to do justice, for example, to article thirteen: the right to freedom of movement. Must the AI be mobile? Does the AI have the right to arms and legs and to all of the other human senses as well?

What about article sixteen: the right to marry and found a family? The Frankenstein monster demanded this right from his creator and was refused. Was Dr. Frankenstein immoral in refusing this request? Does an AI have the right to reproduce by any means? Does "it" have the right to slow growth (that is, a maturation process) or the right to a specific gender?

If we concede a right to freedom from unnecessary pain as well, a right we seem to confer on subpersonal animals (Singer, 1980), we have to face the delicate technical question I alluded to earlier: Is there a way to give pain capacity to an AI without causing it pain, at least no more pain than an ordinary human birth might entail?

Such questions are quite bewildering. Although I have argued that the bottom line of the morality of the experiment *qua* experiment on the subject is whether a consciousness of any quality is better than none at all, and although I have also argued that an unnecessary experiment which carries with it potential for a severely limited existence and, possibly, unnecessary pain, is unethical, I have indicated that if all of the conditions of a sound birth were met *a priori* the experiment might then be considered legitimate.

All this is somewhat silly. Such AIs will probably be built up in layers, and they are not really even babies, yet! Consciousness itself might prove to be an emergent property that might spring forth unexpectedly from a sophisticated level of hardware and software organization, quite as HAL's did in Clarke's *2001: A Space Odyssey.* Then, if we give birth to a Frankenstein monster, it will be too late. This is often the way human beings learn their lessons, and I have no illusions about the tendency of technology to run with its own momentum, to "do it if it can be done." Perhaps, though, if we give free rein to our moral imaginations, we will be better prepared than was poor old Dr. Frankenstein.

Can An Artificial Intelligence Be Moral?

A common criticism of the AI project is that a computer only does what it is programmed to do, that it is without the mysterious property called free will and, therefore, can never become "moral." (I will take free will to mean, specifically, the attribution of an intervening variable between stimulus and response that renders impossible a prediction of response, given adequate knowledge of all input variables.) Some philosophers might even be tempted to say that a machine, however intelligent, which is without the capacity to value and to make moral choices cannot be considered an end in itself and, therefore, could not be the possessor of certain human rights (I think, at least, that this is what Kant's argument would boil down to). Indeed, Kant (1964) would argue that a capacity for free will is a necessary postulate of the moral life, though there is nothing empirical about this capacity. (I cannot enter here into an analysis of Kant's arguments about the ontology of this capacity, for example, that freewill is part of the intelligible world as opposed to the sensual-empirical world. Suffice it to say that I believe this mysterious capacity for free will is really, for Kant, something from another reality and is, thus, subject to the criticism of dualism which I have previously alluded to.)

A cursory glance at the history of theology and philosophy on the topic of free will, from Augustine and Pelagius to Chomsky and Skinner, shows the difficulty of delineating any proof for its existence or nonexistence. For every Chomsky (1973) who maintains that a behaviorist has no predictive ability with regard to human behavior, there is a Skinner who maintains that the behaviorist does not have all the information about input variables necessary to make a complete prediction about behavior or output. Free will and determinism might really be differing perspectives on the same phenomenon, the former being the

perspective from the subjectivity of human agency, the latter from observation. Except for the problem of retributive justice (punishment), I see little or no difference in the "cash value" of holding to one theory or the other; it is interesting to note, however, that the reality of Kant's noumenal capacity for free will might, in a trivial sense, be "proved" by the failure of an AI to ever give behavioristic evidence (however that might be defined) of free will. At any rate, persons make so-called choices to act in certain ways, whether they are factually free or factually determined (or programmed). However, short of invoking the *deus ex machina* of an ontological dualism in order to protect the ghostlike existence of the free will, the contention of its existence really makes little difference to the AI project. If free will is real in some sense, there is again no reason to believe that it might not be an emergent property of a sophisticated level of technical organization, just as it might be asserted to arise through a slow maturation process in humans. I should also add that not all AI experts are convinced an AI could not attain free will (I refer interested persons to the last chapters of Hofstadter's *Gödel, Escher, Bach* for some interesting ruminations on this difficult issue).

Emotion

What about emotion? There has been considerable debate among ethicists about whether the concept of the "good" is a cognitive one or a noncognitive and emotive one (for example, Frankena, 1973). Is morality primarily a matter of reason and logic, or is it a matter of emotion and sympathy? Even for Kant, who was deeply committed to reason, it can be construed to be a matter of both (Paton, 1965). A person, then, must have emotions and reason in order to have the capacity for ethical decision making. Given an accurate, nonreductionistic description of moral experience (Mandelbaum, 1969), an AI would probably have to have feelings and emotions, as well as intelligent reason, in order to replicate personal decision making. The necessity of this capacity becomes more apparent when we seek to discern whether an AI ought to be something like a moral judge. Weizenbaum (1976) has maintained that there are certain things an intelligent machine should not to be allowed to do, especially things involving human emotions such as love. Does this mean, then, that a machine should not be allowed to be a moral judge?

If we took to modern ethical theory in order to ascertain what attributes a competent moral judge must have, we might turn to the ideal observer theory (Firth, 1952). The ideal observer theory maintains that the statement "X is right" means that X would be approved by an ideal moral judge who had following characteristics: omniscience (knowledge of all relevant facts), omnipercipience (the ability to vividly imagine the feelings and circumstances of the parties involved, that is, something like empathy), disinterestedness (nonbiasedness), and dispassionateness (freedom from disturbing passion). These characteristics, then, are an attempt to specify the conditions under which a valid moral judgment might be made. The attributes of such an ideal observer, of course, resemble the traditional attributes of God, and this is understandable if we, like Kant, consider the concept of God to be, in part, a moral ideal. The theory itself, however, does not presuppose a belief in God; it merely contends that it gives an adequate description of the requisite conditions we need in order to make sound moral judgments. We do attempt to approximate these conditions when we make a moral judgment: A judge who shakes hands in the courtroom with your opponent but not with you could justly be accused of failing to be disinterested, and so on. Now if we look at most of the characteristics of such an observer, that is, omniscience, disinterestedness, and dispassionateness, then a case might be made for saying that an unemotional AI could be considered a better moral judge than a human person. Such a machine might be able to store and retrieve more factual data, not be disturbed by violent passions and interests, and so on; it could be said to be capable of "cool" and "detached" choices.

Omnipercipience, or empathy, however, is problematic. This kind of sympathy for the circumstances of people, part of what Aristotle called "equity," is a wisdom that comes from being able to "put ourselves in the other's shoes," as it were. Certainly emotions would be involved here, and the degree of morphological similarity necessary for the empathetic response of one person to another is a subtle and problematic matter. Perhaps the ability to empathize is what Weizenbaum finds lacking in any possible AI and is the reason why he would not entrust such judgments to one. Yet, given what I have previously said there is no reason to believe that such ability could not be, in principle, duplicable.

The Problem of Casuistry

Casuistry deals with the application of general rules to specific, concrete situations. The question of whether all thinking can be formalized in some sort of rule structure is a crucial one for AI in general. For example, one computer program seeks to capture the medical diagnostic ability of a certain physician who has the reputation as one of the best diagnosticians in the world. The computer programmer working with him tries to break this procedure down into a series of logical steps of what to the physician was an irreducible intuition of how to go about doing it. With a lot of prodding, however, the diagnostician was soon able to break these intuitions down into their logical steps (H.E.W., 1980). Perhaps this is true with all "intuitive" thinking, or is it? If we assume that ethics is a reasonable, cognitive undertaking, we are prone to formalize it in a series of rules, not exceptionless rules but something like W. D. Ross's list of *prima facie* obligations: a list of rules, any one of which might be binding in a particular circum-

stance (Ross, 1965). What Ross gives us is something like the moral equivalent of a physicist's periodic table of the elements, moral rules that constitute the elemental building blocks of moral life. The list runs something like this: promise keeping, truth telling, reparations, justice, gratitude, beneficence, nonmaleficence, and self-improvement.

All of these obligations are incumbent upon us as moral beings, but one or several can take precedence over the others in certain situations. We must, therefore, have some principle for adjudicating between these rules in situations where they might conflict. They need not be set up in a strict hierarchy; we could, for example, say that the principle of adjudication is intuition. This is basically what Ross asserts when he quotes Aristotle's famous dictum "The decision rests with the perception." At the same time, however, Ross ranks at least one of his rules as *prima facie* more binding than another. The duty of not harming (nonmaleficence) is a more stringent duty than actively promoting good (beneficence), and this is true even before a particular situation is addressed. This proves significant in what follows.

The problem of the casuistry of an AI has already been imaginatively addressed by Isaac Asimov in his book *I, Robot* (1950), where he lists his somewhat famous "Rules for Robotics." The list is an attempt to give us a rule hierarchy that might be wired into a robot. These rules are as follows:

1. A robot may not injure a human being or through inaction allow a human being to come to harm.
2. A robot must obey orders given it by humans except when such orders conflict with the first law.
3. A robot must protect its own existence as long as such protection does not conflict with the first or second laws

These, of course, are not all of the rules that a robot might be wired with; Ross's list is certainly more complete than Asimov's. There is plenty of food for thought, though, in Asimov's hierarchy.

The first thing an ethicist might notice about rule one is that it attempts to combine the principle of nonmaleficence (do no harm) with a utilitarian principle of beneficence (do the maximum good). Rule one thus contains within itself the potential for conflict, as it does in modern normative theory (Frankena, 1973, pp. 45-48). What if the robot has to decide between injuring a human being or not acting at all and thus allowing another human being to come to harm through an act of omission? Asimov's robots run in circles when confronted with a conflict-of-rule situation. What would be most moral to do in such a situation—refrain from acting at all, or heed the voice of Kierkegaard and act whatever the costs might be? Further, how do we understand the principle of beneficence underlying the sins of omission? Should all robots go to Pakistan in order to maximize the good, or should they go somewhere else? How long should a robot calculate potential consequences before acting? I should point out that if there is any specific normative theory attributed to the AIs of science fiction, it would have to be utilitarian. Robots are seen as paradigms of calculation, as exhibiting metahuman capacities for weighing, quantifying, and projecting consequences. As such, they are subject to the same criticisms one might level at utilitarians in general; for example, how might a robot compare incommensurable goods in order that they might be quantified and rendered mathematically precise? Such a problem vexed Jeremy Bentham for most of his life—is push pin really as good as poetry?

The second rule, obeying orders for humans except where they might conflict with the first rule, appears to contradict the aforementioned right to freedom from slavery (unless, of course, the AI were to be somehow considered a "child"). The second part of the rule, refusing to take an order to harm, might not be considered moral at all if, for example, the protection of innocence were at stake. Should an AI robot be a proponent of the just war theory, or should it be something of a pacifist, even to the point of self-sacrifice as Asimov's robots would appear to be?

The third rule, protecting its own existence as long as such protection does not conflict with the first and second laws, is also highly problematic. First, it gives the robot the right to self-defense but then takes it away. Note the use of the word must. There is no possibility for what ethics calls "supererogatory" duties. These are the things we think it is praiseworthy to do, but not blameworthy not to do. (Urmson, 1958; Chisolm, 1963). Suppose, for example, that a man jumps on a hand grenade in order to save his comrades. His heroic action is likely to be praised, but had he chosen not to jump he probably would not be blamed for failing to sacrifice himself. We like to keep the heroic free from the obligatory. Asimov's rules do not. Is that really bad? Should we wire a self-sacrificial attitude into our robots, making them all little Christs? For that matter, should we "wire" our own children to so act? The questions involved in wiring a robot for morality are so very similar to the questions of how we should morally educate our own children!

It might prove to be the case that no hierarchy of normative principles can do justice to the complexity of personal, moral choice. It also might be that the self-reflexively conscious ego of a sophisticated AI would take no programming at all, and that it would pick and choose its own rules, rules it learns through the trials and errors of time. However difficult it might prove to duplicate human, moral decision making, especially an adjudicative principle like intuition, we need not resort to a skepticism that is based ultimately on dualistic "magic," and thereby resign from the attempt.

Conclusion

What if we never make an AI that can pass the Turing Test? Little of the effort will be lost. It is the peculiar pedagogical effect of "distancing" that makes the contemplation of artificial persons so fertile for the human imagination. The proponents of AI promise us that we will learn more about ourselves in the attempt to construct something like ourselves. This distancing is also dangerous, however. We have, for example, distanced ourselves from other people and from animals, often with tragic results. Of course, the project will go on and, I think, with much success, but it will be a sad thing if Hegel was right when he said, "The owl of Minerva flies only at midnight." Much caution and forethought are necessary when we contemplate the human construction of the personal.

If we can't ever make such an intelligence, is any mystery gone? To the contrary, the failure to be able to produce the personal as artifact might eventually bring us to the brink of a mysticism that has, at least, been partially "tested." Would it be more mysterious to find intelligent life elsewhere in the universe or to find after unimaginable aeons that we are unique and alone? Perhaps AI is the next stage of evolution, a harder blow to our ineradicable anthropomorphism than Copernicus's theory that the planets revolved around the sun and not the earth. As McCorduck says of the undertaking, "Face to face with mind as artifact we're face to face with almost more themes in the human experience than we can count or comprehend. And there's the added zest that this idea may turn out to transcend the human experience altogether and lead us to the metahuman" (p. 329).

On one side of the moral spectrum lies the fear of something going wrong; on the other side is the exuberant "yes" to all possibilities and benefits. Though the first word of ethics is "do no harm," we can perhaps look forward to innovation with a thoughtful caution. Perhaps we will eclipse ourselves with our own inventions. Perhaps Michelangelo was correct when he pointed that long finger of God at Adam's hand. Either way, I am excited.

References

Asimov, I (1950) *I robot.* New York: Gnome Press

Boden, M (1977) *Artificial intelligence and natural man.* New York: Basic Books.

Boeyink, David (1974) Pain and suffering *Journal of Religious Ethics* 2(1): 85-97.

Capek, The Brothers (1975). *R U.R and the insect play.* London: Oxford University Press

Chisholm, Roderick (1963) Supererogation and offense: A conceptual scheme for ethics *Ratio* 5(June): 1-14

Chomsky, Noam (1973) *For reasons of state.* New York: Random House

Department of Health, Education, & Welfare (1980) *The seeds of artificial intelligence: Sumex-aim.* Washington, D C.

Dreyfus, H (1972). *What computers can't do: A critique of artificial reason.* New York: Harper & Row

Dyck, A. J (1977) Ethics and medicine In S. J Reiser, A. J. Dyck, & W. J. Curran (Eds) *Ethics in medicine: Historical perspectives and contemporary concerns.* Cambridge, Mass: MIT Press

Firth, Roderick (1952) Ethical absolutism and the ideal observer *Philosophy and Phenomenological Research* 12 (March): 317-345

Fletcher, Joseph (1977) Ethical aspects of genetic controls. In S J Reiser, A J Dyck, & W J. Curran (Eds) *Ethics in medicine: Historical perspectives and contemporary concerns.* Cambridge, Mass.: MIT Press

Fletcher, Joseph (1972). Indicators of humanhood: A tentative profile of man. *The Hastings Center Report* 2(5):1-4

Frankena, W K (1973) *Ethics.* Englewood Cliffs, N J : Prentice-Hall

Heidegger, Martin (1962) *Being and time* New York: Harper & Row.

Hofstadter, Douglas R. (1980) *Gödel, Escher, Bach: An eternal golden braid.* New York: Vintage Books.

Jonas, Hans (1977) Philosophical reflections on experimenting with human subjects. In S. J Reiser, A. J Dyck, & W J. Curran (Eds) *Ethics in medicine: Historical perspectives and contemporary concerns.* Cambridge, Mass.: MIT Press

Kant, Immanuel (1964) *Groundwork of the metaphysic of morals.* New York: Harper & Row

McCorduck, Pamela (1979) *Machines who think* San Francisco: W H. Freeman

Mandelbaum, Maurice (1969) *The phenomenology of moral experience.* Baltimore, Md : Johns Hopkins Press

Marx, Karl, & Engels, F. (1971) *On religion* New York: Schocken

Mead, G H. (1972) *Mind, self, and society* Chicago: University of Chicago Press

Niebuhr, H R (1963) *The responsible self.* New York: Harper & Row

Noonan, John T (1970) An almost absolute value in history In J T Noonan (Ed.) *The morality of abortion* Cambridge: Harvard University Press

Paton, H. J (1965). *The categorical imperative.* New York: Harper & Row.

Penfield, Wilder (1983). The uncommitted cortex: The child's changing brain *The Atlantic Monthly* 22(7): 77-81

Ramsey, Paul (1975) *The ethics of fetal research* New Haven, Conn : Yale University Press.

Roe vs Wade, decision on abortion. (1977) In S J. Reiser, A. J Dyck, & W J Curran (Eds) *Ethics in medicine: Historical perspectives and contemporary concerns* Cambridge, Mass.: MIT Press.

Rosenthal, Elisabeth (1983) A rebel in the computer revolution *Science Digest* (August): 94-97.

Ross, W D (1965) *The right and the good* Oxford: Clarendon

Sartre, Jean-Paul (1957) *Existentialism and human emotions.* New York: Wisdom Library.

Schlein, J M (1961) A client-centered approach to schizophrenia In A Burton (Ed) *Psychotherapy of the psychoses.* New York: Basic Books

Shelley, Mary W (1974) *Frankenstein* New York: Scholastic Book Services.

Singer, Peter (1980). *Practical ethics* New York: Cambridge University Press

Tooley, Michael (1974) Abortion and infanticide In Cohen, Nagel, & Scanlon (Eds) *The rights and wrongs of abortion.* Princeton, N J : Princeton University Press

United Nations (1979) The universal declaration of human rights, 1948 In T Donaldson & P Werhane (Eds) *Ethical issues in business* Englewood Cliffs, N J.: Prentice-Hall

Urmson, J. O (1958) *Saints and heroes.* Seattle: University of Washington Press

Weizenbaum, Joseph (1976) *Computer power and human reason* San Francisco: W H. Freeman

Wiener, Norbert (1964) *God and golem, inc* Cambridge, Mass : MIT Press.