AN UNREASONABLE BOOK

1976

Joseph Weizenbaum, *Computer Power and Human Reason*, W. H. Freeman Co., San Francisco 1975

This moralistic and incoherent book uses computer science and technology as an illustration to support the view promoted by Lewis Mumford, Theodore Roszak, and Jacques Ellul, that science has led to an immoral view of man and the world. I am frightened by its arguments that certain research should not be done if it is based on or might result in an "obscene" picture of the world and man. Worse yet, the book's notion of "obscenity" is vague enough to admit arbitrary interpretations by activist bureaucrats.

1 It's Hard to Figure Out What He Really Believes . . .

Weizenbaum's style involves making extreme statements which are later qualified by contradictory statements. Therefore, almost any quotation is out of context, making it difficult to summarize his contentions accurately.

The following passages illustrate the difficulty:

"In 1935, Michael Polanyi", [British chemist and philosopher of science, was told by] *"Nicolai Bukharin, one of the leading theoreticians of the Russian Communist party, . . . [that] 'under socialism the conception of science pursued for its own sake would disappear, for the interests of scientists would spontaneously turn to the problems of the current Five Year Plan.' Polanyi sensed then that 'the scientific outlook appeared to have produced a mechanical conception of man and history in which there was no place for science itself.' And further that 'this conception denied altogether any intrinsic power to thought and thus denied any grounds for claiming freedom of thought.' "*

- from page 1. Well, that's clear enough; Weizenbaum favors freedom of thought and science and is worried about threats to them. But on page 265, we have

"Scientists who continue to prattle on about 'knowledge for its own sake' in order to exploit that slogan for their self-serving ends have detached science and knowledge from any contact with the real world". Here Weizenbaum seems to be against pure science, i.e. research motivated solely by curiosity. We also have

"With few exceptions, there have been no results, from over twenty years of artificial intelligence research, that have found their way into industry generally or into the computer industry in particular". - page 229 This again suggests that industrial results are necessary to validate science.

"Science promised man power. But as so often happens when people are seduced by promises of power ... the price actually paid is servitude and impotence". This is from the book jacket. Presumably the publisher regards it as a good summary of the book's main point.

"I will, in what follows, try to maintain the position that there is nothing wrong with viewing man as an information processor (or indeed as anything else) nor with attempting to understand him from that perspective, providing, however, that we never act as though any single perspective can comprehend the whole man." - page 140. We can certainly live with that, but

"Not only has our unbounded feeding on science caused us to become dependent on it, but, as happens with many other drugs taken in increasing dosages, science has been gradually converted into a slow acting poison". - page 13. These are qualified by

"I argue for the rational use of science and technology, not for its mystification, let alone its abandonment". - page 256

In reference to the proposal for a moratorium on certain experiments with recombinant DNA because they might be dangerous, we have *"Theirs is certainly a step in the right direction, and their initiative is to be applauded. Still, one may ask, why do they feel they have to give a reason for what they recommend at all? Is not the overriding obligation on men, including men of science, to exempt life itself from the madness of treating everything as an object, a sufficient reason, and one that does not even have to be spoken? Why does it have to be explained? It would appear that even the noblest acts of the most well-meaning people are poisoned by the corrosive climate of values of our time."* Is Weizenbaum against all experimental biology or even all experiments with DNA? I would hesitate to conclude so from this quote; he may say the direct opposite somewhere else. Weizenbaum's goal of getting lines of research abandoned without even having to give a reason seems unlikely to be achieved except in an atmosphere that combines public hysteria and bureaucratic power. This has happened under conditions of religious enthusiasm and in Nazi Germany, in Stalinist Russia and in the China of the "Cultural Revolution". Most likely it won't happen in America.

"Those who know who and what they are do not need to ask what they should do." - page 273. Let me assure the reader that there is nothing in the book that offers any way to interpret this pomposity. I take it as another plea to be free of the bondage of having to give reasons for his denunciations.

The menace of such grandiloquent precepts is that they require a priesthood to apply them to particular cases, and would-be priests quickly crystallize around any potential center of power. A corollary of this is that people can be attacked for what they are rather than for anything specific they have done. The April 1976 issue of *Ms.* has a poignant illustration of this in an article about "trashing".

"An individual is dehumanized whenever he is treated as less than a whole person". - page 266. This is also subject to priestly interpretation as in the encounter group movement.

"The first kind" [of computer application] *"I would call simply obscene. These are ones whose very contemplation ought to give rise to feelings of disgust in every civilized person. The proposal I have mentioned, that an animal's visual system and brain be coupled to computers, is an example. It represents an attack on life itself. One must wonder what must have happened to the proposers' perception of life, hence to their perceptions of themselves as part of the continuum of life, that they can even think of such a thing, let alone advocated it"*. No argument is offered that might be answered, and no attempt is made to define criteria of acceptability. I think Weizenbaum and the scientists who have praised the book may be surprised at some of the repressive uses to which the book will be put. However, they will be able to point to passages in the book with quite contrary sentiments, so the repression won't be their fault.

2 But Here's a Try at Summarizing

As these inconsistent passages show, it isn't easy to determine Weizenbaum's position, but the following seem to be the book's main points:

1. Weizenbaum doesn't name any specific task that computers cannot carry out, because he wishes *"to avoid the unnecessary, interminable, and ultimately sterile exercise of making a catalogue of what computers will and will not be able to do, either here and now or ever"*. It is also stated that human and machine reasoning are incomparable and that the sensory experience of a human is essential for human reasoning.

2. There are tasks that computers should not be programmed to do.

   Some are tasks Weizenbaum thinks shouldn't be done at all - perhaps for political reasons. One may quarrel with his politics, and I do, but obviously computers shouldn't do what shouldn't be done. However, Weizenbaum also objects to computer hookups to animal brains and computer conducted psychiatric interviews. As to the former, I couldn't tell whether he is an anti-vivisectionist, but he seems to have additional reasons for calling them "obscene". The objection to computers doing psychiatric interviews also has a component beyond the conviction that they would necessarily do it badly. Thus he says, *"What can the psychiatrist's image of his patient be when he sees himself, as a therapist, not as an engaged human being acting as a healer, but as an information processor following rules, etc.?"* This seems like the renaissance era religious objections to dissecting the human body that came up when science revived. Even the Popes eventually convinced themselves that regarding the body as a machine for scientific or medical purposes was quite compatible with regarding it as the temple of the soul. Recently they have taken the same view of studying mental mechanisms for scientific or psychiatric purposes.

3. Science has led people to a wrong view of the world and of life.

   The view is characterized as mechanistic, and the example of clockwork is given. (It seems strange for a computer scientist to give this example, because the advance of the computer model over older mechanistic models is that computers can and clockwork can't make decisions.) Apparently analysis of a living system as composed of interacting parts rather than treating it as an unanalyzed whole is bad.

4. Science is not the sole or even main source of reliable general knowledge.

   However, he doesn't propose any other sources of knowledge or say what the limits of scientific knowledge is except to characterize certain thoughts as “obscene”.

5. Certain people and institutions are attacked.

   These include the Department of “Defense” (sic), *Psychology Today, The New York Times* Data Bank, compulsive computer programmers, Kenneth Colby, Marvin Minsky, Roger Schank, Allen Newell, Herbert Simon, J.W. Forrester, Edward Fredkin, B.F. Skinner, Warren McCulloch (until he was old), Laplace and Leibniz.

6. Certain political and social views are taken for granted.

   The view that U.S. policy in Vietnam was “murderous” is used to support an attack on “logicality” (as opposed to “rationality”) and the view of science as a ”slow acting poison”. The phrase *“It may be that the people’s cultivated and finally addictive hunger for private automobiles . . .”* (p.30) makes psychological, sociological, political, and technological presumptions all in one phrase. Similarly, *“Men could instead choose to have truly safe automobiles, decent television, decent housing for everyone, or comfortable, safe, and widely distributed mass transportation.”* presumes wide agreement about what these things are, what is technologically feasible, what the effects of changed policies would be, and what activities aimed at changing people’s taste are permissible for governments.

3 The ELIZA Example

Perhaps the most interesting part of the book is the account of his own program ELIZA that parodies Rogerian non-directive psychotherapy and his anecdotal account of how some people ascribe intelligence and personality to it. In my opinion, it is quite natural for people who don’t understand the notion of algorithm to imagine that a computer computes analogously to the way a human reasons. This leads to the idea that accurate computation entails correct reasoning and even to the idea that computer malfunctions are analogous to human neuroses and psychoses. Actually, programming a computer to draw interesting conclusions from premises is very difficult and only limited success has been attained. However, the effect of these natural misconceptions shouldn’t be exaggerated; people readily understand the truth when it is explained, especially when it applies to a matter that concerns them. In particular, when an executive excuses a mistake by saying that he placed excessive faith in a computer, a certain skepticism is called for.

Colby's (1973) study is interesting in this connection, but the interpretation below is mine. Colby had psychiatrists interview patients over a teletype line and also had them interview his PARRY program that simulates a paranoid. Other psychiatrists were asked to decide from the transcripts whether the interview was with a man or with a program, and they did no better than chance. However, since PARRY is incapable of the simplest causal reasoning, if you ask, "How do you know the people following you are Mafia" and get a reply that they look like Italians, this must be a man not PARRY. Curiously, it is easier to imitate (well enough to fool a psychiatrist) the emotional side of a man than his intellectual side. Probably the subjects expected the machine to have more logical ability, and this expectation contributed to their mistakes. Alas, random selection from the directory of the Association for Computing Machinery did no better.

It seems to me that ELIZA and PARRY show only that people, including psychiatrists, often have to draw conclusions on slight evidence, and are therefore easily fooled. If I am right, two sentences of instruction would allow them to do better.

In his 1966 paper on ELIZA (cited as 1965), Weizenbaum writes,

"One goal for an augmented ELIZA program is thus a system which already has access to a store of information about some aspect of the real world and which, by means of conversational interaction with people, can reveal both what it knows, i.e. behave as an information retrieval system, and where its knowledge ends and needs to be augmented. Hopefully the augmentation of its knowledge will also be a direct consequence of its conversational experience. It is precisely the prospect that such a program will converse with many people and learn something from each of them which leads to the hope that it will prove an interesting and even useful conversational partner." Too bad he didn't successfully pursue this goal; no-one else has. I think success would have required a better understanding of formalization than is exhibited in the book.

4 What Does He Say About Computers?

While Weizenbaum's main conclusions concern science in general and are moralistic in character, some of his remarks about computer science and AI are worthy of comment.

1. He concludes that since a computer cannot have the experience of a man, it cannot understand a man. There are three points to be made in reply. First, humans share each other's experiences and those of machines or animals only to a limited extent. In particular, men and women have different experiences. Nevertheless, it is common in literature for a good writer to show greater understanding of the experience of the opposite sex than a poorer writer of that sex. Second, the notion of experience is poorly understood; if we understood it better, we could reason about whether a machine could have a simulated or vicarious experience normally confined to humans. Third, what we mean by understanding is poorly understood, so we don't yet know how to define whether a machine understands something or not.

2. Like his predecessor critics of artificial intelligence, Taube, Dreyfus and Lighthill, Weizenbaum is impatient, implying that if the problem hasn't been solved in twenty years, it is time to give up. Genetics took about a century to go from Mendel to the genetic code for proteins, and still has a long way to go before we will fully understand the genetics and evolution of intelligence and behavior. Artificial intelligence may be just as difficult. My current answer to the question of when machines will reach human-level intelligence is that a precise calculation shows that we are between 1.7 and 3.1 Einsteins and .3 Manhattan Projects away from the goal. However, the current research is producing the information on which the Einstein will base himself and is producing useful capabilities all the time.

3. The book confuses computer simulation of a phenomenon with its formalization in logic. A simulation is only one kind of formalization and not often the most useful - even to a computer. In the first place, logical and mathematical formalizations can use partial information about a system insufficient for a simulation. Thus the law of conservation of energy tells us much about possible energy conversion systems before we define even one of them. Even when a simulation program is

available, other formalizations are necessary even to make good use of the simulation. This review isn't the place for a full explanation of the relations between these concepts.

Like *Punch*'s famous curate's egg, the book is good in parts. Thus it raises the following interesting issues:

1. What would it mean for a computer to hope or be desperate for love? Answers to these questions depend on being able to formalize (not simulate) the phenomena in question. My guess is that adding a notion of hope to an axiomatization of belief and wanting might not be difficult. The study of *propositional attitudes* in philosophical logic points in that direction.

2. Do differences in experience make human and machine intelligence necessarily so different that it is meaningless to ask whether a machine can be more intelligent than a machine? My opinion is that comparison will turn out to be meaningful. After all, most people have no doubt that humans are more intelligent than turkeys. Weizenbaum's examples of the dependence of human intelligence on sensory abilities seem even refutable, because we recognize no fundamental difference in humanness in people who are severely handicapped sensorily, e.g. the deaf, dumb and blind or paraplegics.

5 In Defense of the Unjustly Attacked—Some of whom are Innocent

Here are defenses of Weizenbaum's targets. They are not guaranteed to entirely suit the defendees.

Weizenbaum's conjecture that the Defense Department supports speech recognition research in order to be able to snoop on telephone conversations is biased, baseless, false, and seems motivated by political malice. The committee of scientists that proposed the project advanced quite different considerations, and the high officials who made the final decisions are not ogres. Anyway their other responsibilities leave them no time for complicated and devious considerations. I put this one first, because I think the failure of many scientists to defend the Defense Department against attacks they know are unjustified, is unjust in itself, and furthermore has harmed the country.

Weizenbaum doubts that computer speech recognition will have cost-effective applications beyond snooping on phone conversations. He also says, *"There is no question in my mind that there is no pressing human problem that will be more easily solved because such machines exist"*. I worry more about whether the programs can be made to work before the sponsor loses patience. Once they work, costs will come down. Winograd pointed out to me that many possible household applications of computers may not be feasible without some computer speech recognition. One needs to think **both** about how to solve recognized problems technological possibilities to good use. The telephone was not invented by a committee considering already identified problems of communication.

Referring to *Psychology Today* as a cafeteria simply excites the snobbery of those who would like to consider their psychological knowledge to be above the popular level. So far as I know, professional and academic psychologists welcome the opportunity offered by *Psychology Today* to explain their ideas to a wide public. They might even buy a cut-down version of Weizenbaum's book if he asks them nicely. Hmm, they might even buy this review.

Weizenbaum has invented a *New York Times Data Bank* different from the one operated by *The New York Times* - and possibly better. The real one stores abstracts written by humans and doesn't use the tapes intended for typesetting machines. As a result the user has access only to abstracts and cannot search on features of the stories themselves, i.e. he is at the mercy of what the abstractors thought was important at the time.

Using computer programs as psychotherapists, as Colby proposed, would be moral if it would cure people. Unfortunately, computer science isn't up to it, and maybe the psychiatrists aren't either.

I agree with Minsky in criticizing the reluctance of art theorists to develop formal theories. George Birkhoff's formal theory was probably wrong, but he shouldn't have been criticized for trying. The problem seems very difficult to me, and I have made no significant progress in responding to a challenge from Arthur Koestler to tell how a computer program might make or even recognize jokes. Perhaps some reader of this review might have more success.

There is a whole chapter attacking "compulsive computer programmers" or "hackers". This mythical beast lives in the computer laboratory, is an expert on all the ins and outs of the time-sharing system, elaborates the time-sharing system with arcane features that he never documents, and is always changing the system before he even fixes the bugs in the previous version. All these vices exist, but I can't think of any individual who combines them, and people generally outgrow them. As a laboratory director, I have to protect the interests of people who program only part time against tendencies to over-complicate the facilities. People who spend all their time programming and who exchange information by word of mouth sometimes have to be pressed to make proper writeups. The other side of the issue is that we professors of computer science sometimes lose our ability to write actual computer programs through lack of practice and envy younger people who can spend full time in the laboratory. The phenomenon is well known in other sciences and in other human activities.

Weizenbaum attacks the Yale computer linguist, Roger Schank, as follows - the inner quotes are from Schank: *"What is contributed when it is asserted that 'there exists a conceptual base that is interlingual, onto which linguistic structures in a given language map during the understanding process and out of which such structures are created during generation [of linguistic utterances]'? Nothing at all. For the term 'conceptual base' could perfectly well be replaced by the word 'something'. And who could argue with that so-transformed statement?"* Weizenbaum goes on to say that the real scientific problem "remains as untouched as ever". On the next page he says that unless the "Schank-like scheme" understood the sentence *"Will you come to dinner with me this evening?"* to mean *"a shy young man's desperate longing for love"*, then the sense in which the system "understands" is "about as weak as the sense in which ELIZA "understood". This good example raises interesting issues and seems to call for some distinctions. Full understanding of the sentence indeed results in knowing about the young man's desire for love, but it would seem that there is a useful lesser level of understanding in which the machine would know only that he would like her to come to dinner.[1]

Contrast Weizenbaum's demanding, more-human-than-thou attitude to Schank and Winograd with his respectful and even obsequious attitude to Chomsky. We have *"The linguist's first task is therefore to write grammars, that is, sets of rules, of particular languages, grammars capable of characterizing all and only the grammatically admissible sentences of those languages, and then to postulate principles from which crucial features of all such gram-*

[1] 2000 note: That's full understanding in context. The lesser understanding is far beyond what Eliza-like methods can do.

mars can be deduced. That set of principles would then constitute a universal grammar. Chomsky's hypothesis is, to put it another way, that the rules of such a universal grammar would constitute a kind of projective description of important aspects of the human mind." There is nothing here demanding that the universal grammar take into account the young man's desire for love. As far as I can see, Chomsky is just as much a rationalist as we artificial intelligentsia.

Chomsky's goal of a universal grammar and Schank's goal of a conceptual base are similar, except that Schank's ideas are further developed, and the performance of his students' programs can be compared with reality. I think they will require drastic revision and may not be on the right track at all, but then I am pursuing a rather different line of research concerning how to represent the basic facts that an intelligent being must know about the world. My idea is to start from epistemology rather than from language, regarding their linguistic representation as secondary. This approach has proved difficult, has attracted few practitioners, and has led to few computer programs, but I still think it's right.

Weizenbaum approves of the Chomsky school's haughty attitude towards Schank, Winograd and other AI based language researchers. On page 184, he states, *"many linguists, for example, Noam Chomsky, believe that enough thinking about language remains to be done to occupy them usefully for yet a little while, and that any effort to convert their present theories into computer models would, if attempted by the people best qualified, be a diversion from the main task. And they rightly see no point to spending any of their energies studying the work of the hackers."*

This brings the chapter on "compulsive computer programmers" alias "hackers" into a sharper focus. Chomsky's latest book *Reflections on Language* makes no reference to the work of Winograd, Schank, Charniak, Wilks, Bobrow or William Woods to name only a few of those who have developed large computer systems that work with natural language and who write papers on the semantics of natural language. The actual young computer programmers who call themselves hackers and who come closest to meeting Weizenbaum's description don't write papers on natural language. So it seems that the hackers whose work need not be studied are Winograd, Schank, et. al. who are professors and senior scientists. The Chomsky school may be embarassed by the fact that it has only recently arrived at the conclusion that the semantics of natural language is more fundamental than its syntax, while AI based researchers have been pursuing this line for fifteen years.

The outside observer should be aware that to some extent this is a pillow fight within M.I.T. Chomsky and Halle are not to be dislodged from M.I.T. and neither is Minsky - whose students have pioneered the AI approach to natural language. Schank is quite secure at Yale. Weizenbaum also has tenure. However, some assistant professorships in linguistics may be at stake, especially at M.I.T.

Allen Newell and Herbert Simon are criticized for being overoptimistic and are considered morally defective for attempting to describe humans as difference-reducing machines. Simon's view that the human is a simple system in a complex environment is singled out for attack. In my opinion, they were overoptimistic, because their GPS model on which they put their bets wasn't good enough. Maybe Newell's current *production system models* will work out better. As to whether human mental structure will eventually turn out to be simple, I vacillate but incline to the view that it will turn out to be one of the most complex biological phenomena.

I regard Forrester's models as incapable of taking into account qualitative changes, and the world models they have built as defective even in their own terms, because they leave out saturation-of-demand effects that cannot be discovered by curve-fitting as long as a system is only rate-of-expansion limited. Moreover, I don't accept his claim that his models are better suited than the unaided mind in "interpreting how social systems behave", but Weizenbaum's sarcasm on page 246 is unconvincing. He quotes Forrester, "[desirable modes of behavior of the social system] *seem to be possible only if we have a good understanding of the system dynamics and are willing to endure the self-discipline and pressures that must accompany the desirable mode' "*. Weizenbaum comments, *"There is undoubtedly some interpretation of the words 'system' and 'dynamics' which would lend a benign meaning to this observation"*. Sorry, but it looks ok to me provided one is suitably critical of Forrester's proposed social goals and the possibility of making the necessary assumptions and putting them into his models.

Skinner's behaviorism that refuses to assign reality to people's internal state seems wrong to me, but we can't call him immoral for trying to convince us of what he thinks is true.

Weizenbaum quotes Edward Fredkin, former director of Project MAC, and the late Warren McCulloch of M.I.T. without giving their names. pp. 241 and 240. Perhaps he thinks a few puzzles will make the book more interesting, and this is so. Fredkin's plea for research in automatic programming seems to overestimate the extent to which our society currently relies on computers for decisions. It also overestimates the ability of the faculty of a particular university to control the uses to which technology will be put, and it underestimates the difficulty of making knowledge based systems of practical use. Weizenbaum is correct in pointing out that Fredkin doesn't mention the existence of genuine conflicts in society, but only the new left sloganeering elsewhere in the book gives a hint as to what he thinks they are and how he proposes to resolve them.

As for the quotation from (McCulloch 1956), Minsky tells me "this is a brave attempt to find a dignified sense of freedom within the psychological determinism morass". Probably this can be done better now, but Weizenbaum wrongly implies that McCulloch's 1956 effort is to his moral discredit.

Finally, Weizenbaum attributes to me two statements - both from oral presentations - which I cannot verify. One of them is *"The only reason we have not yet succeeded in simulating every aspect of the real world is that we have been lacking a sufficiently powerful logical calculus. I am working on that problem"*. This statement doesn't express my present opinion or my opinion in 1973 when I am alleged to have expressed it in a debate, and no-one has been able to find it in the video-tape of the debate.

We can't simulate "every aspect of the real world", because the initial state information is never available, the laws of motion are imperfectly known, and the calculations for a simulation are too extensive. Moreover, simulation wouldn't necessarily answer our questions. Instead, we must find out how to represent in the memory of a computer the information about the real world that is actually available to a machine or organism with given sensory capability, and also how to represent a means of drawing those useful conclusions about the effects of courses of action that can be correctly inferred from the attainable information. Having *a sufficiently powerful logical calculus* is an important part of this problem—but one of the easier parts.

[**Note added September 1976** - This statement has been quoted in a large fraction of the reviews of Weizenbaum's book (e.g. in *Datamation* and *Nature*) as an example of the arrogance of the "artificial intelligentsia". Weizenbaum firmly insisted that he heard it in the Lighthill debate and cited his notes as corroboration, but later admitted (in *Datamation*) after reviewing the tape that he didn't, but claimed I must have said it in some other debate. I am confident I didn't say it, because it contradicts views I have held and repeatedly stated since 1959. My present conjecture is that Weizenbaum heard me say something on the importance of formalization, couldn't quite remember what, and quoted ”what McCarthy must have said” based on his own misunderstanding of the relation between computer modeling and formalization. (His two chapters on computers show no awareness of the difference between declarative and procedural knowledge or of the discussions in the AI literature of their respective roles). Needless to say, the repeated citation by reviewers of a pompous statement that I never made and which is in opposition to the view that I think represents my major contribution to AI - is very offensive].

The second quotation from me is the rhetorical question, *“What do judges know that we cannot tell a computer”*. I’ll stand on that if we make it “eventually tell” and especially if we require that it be something that one human can reliably teach another.

6 A Summary of Polemical Sins

The speculative sections of the book contain numerous dubious little theories, such as this one about the dehumanizing effect of of the invention of the clock: *“The clock had created literally a new reality; and that is what I meant when I said earlier that the trick man turned that prepared the scene for the rise of modern science was nothing less than the transformation of nature and of his perception of reality. It is important to realize that this newly created reality was and remains an impoverished version of the older one, for it rests on a rejection of those direct experiences that formed the basis for, and indeed constituted the old reality. The feeling of hunger was rejected as a stimulus for eating; instead one ate when an abstract model had achieved a certain state, i.e. when the hand of a clock pointed to certain marks on the clock’s face (the anthropomorphism here is highly significant too), and similarly for signals for sleep and rising, and so on.”*

This idealization of primitive life is simply thoughtless. Like modern man, primitive man ate when the food was ready, and primitive man probably had to start preparing it even further in advance. Like modern man, primitive man lived in families whose members are no more likely to become hungry all at once than are the members of a present family.

I get the feeling that in toppling this microtheory I am not playing the game; the theory is intended only to provide an atmosphere, and like the reader of a novel, I am supposed to suspend disbelief. But the contention that science has driven us from a psychological Garden of Eden depends heavily on such word pictures.

By the way, I recall from my last sabbatical at M.I.T. that the *feeling of hunger* is more often *the direct social stimulus for eating* for the "hackers" deplored in Chapter 4 than it could have been for primitive man. Often on a crisp New England night, even as the clock strikes three, I hear them call to one another, messages flash on the screens, a flock of hackers magically gathers, and the whole picturesque assembly rushes chattering off to Chinatown.

I find the book substandard as a piece of polemical writing in the following respects:

1. The author has failed to work out his own positions on the issues he discusses. Making an extreme statement in one place and a contradictory statement in another is no substitute for trying to take all the factors into account and reach a considered position. Unsuspicious readers can come away with a great variety of views, and the book can be used to support contradictory positions.

2. The computer linguists - Winograd, Schank, et. al. - are denigrated as hackers and compulsive computer programmers by innuendo.

3. One would like to know more precisely what biological and psychological experiments and computer applications he finds acceptable. Reviewers have already drawn a variety of conclusions on this point.

4. The terms "authentic", "obscene", and "dehumanization" are used as clubs. This is what mathematicians call "proof by intimidation".

5. The book encourages a snobbery that has no need to argue for its point of view but merely utters code words, on hearing which the audience is supposed applaud or hiss as the case may be. The *New Scientist* reviewer certainly salivates in most of the intended places.

6. Finally, when moralizing is both vehement and vague, it invites authoritarian abuse either by existing authority or by new political movements. Imagine, if you can, that this book were the bible of some bureaucracy, e.g. an Office of Technology Assessment, that acquired power over the computing or scientific activities of a university, state, or country. Suppose Weizenbaum's slogans were combined with *the bureaucratic ethic*

> that holds that any problem can be solved by a law forbidding something and a bureaucracy of eager young lawyers to enforce it. Postulate further a vague *Humane Research Act* and a "public interest" organization with more eager young lawyers suing to get judges to legislate new interpretations of the Act. One can see a laboratory needing more lawyers than scientists and a Humane Research Administrator capable of forbidding or requiring almost anything.

I see no evidence that Weizenbaum forsees his work being used in this way; he doesn't use the phrase *laissez innover* which is the would-be science bureaucrat's analogue of the economist's *laissez faire*, and he never uses the indefinite phrase "it should be decided" which is a common expression of the bureaucratic ethic. However, he has certainly given his fellow computer scientists at least some reason to worry about potential tyranny.

Let me conclude this section with a quotation from Andrew D. White, the first president of Cornell University, that seems applicable to the present situation - not only in computer science, but also in biology. - *"In all modern history, interference with science in the supposed interest of religion, no matter how conscientious such interference may have been, has resulted in the direst evils both to religion and to science, and invariably; and, on the other hand, all untrammelled scientific investigation, no matter how dangerous to religion some of its stages my have seemed for the time to be, has invariably resulted in the highest good both of religion and of science"*. Substitute *morality* for *religion* and the parallel is clear. Frankly, the feebleness of the reaction to attacks on scientific freedom worries me more than the strength of the attacks.

7 What Worries about Computers are Warranted?

Grumbling about Weizenbaum's mistakes and moralizing is not enough. Genuine worries prompted the book, and many people share them. Here are the genuine concerns that I can identify and the opinions of one computer scientist about their resolution: What is the danger that the computer will lead to a false model of man? What is the danger that computers will be misused? Can human-level artificial intelligence be achieved? What, if any, motivational characteristics will it have? Would the achievement of artificial intelligence be good or bad for humanity?

1. **Does the computer model lead to a false model of man.**

   Historically, the mechanistic model of the life and the world followed animistic models in accordance with which, priests and medicine men tried to correct malfunctions of the environment and man by inducing spirits to behave better. Replacing them by mechanistic models replaced shamanism by medicine. Roszak explicitly would like to bring these models back, because he finds them more "human", but he ignores the sad fact that they don't work, because the world isn't constructed that way. The pre-computer mechanistic models of the mind were, in my opinion, unsuccessful,but I think the psychologists pursuing computational models of mental processes may eventually develop a really beneficial psychiatry.

   Philosophical and moral thinking hasn't yet found a model of man that relates human beliefs and purposes to the physical world in a plausible way. Some of the unsuccessful attempts have been more mechanistic than others. Both mechanistic and non-mechanistic models have led to great harm when made the basis of political ideology, because they have allowed tortuous reasoning to justify actions that simple human intuition regards as immoral. In my opinion, the relation between beliefs, purposes and wants to the physical world is a complicated but ultimately solvable problem. Computer models can help solve it, and can provide criteria that will enable us to reject false solutions. The latter is more important for now, and computer models are already hastening the decay of dialectical materialism in the Soviet Union.

2. **What is the danger that computers will be misused?**

   Up to now, computers have been just another labor-saving technology. I don't agree with Weizenbaum's acceptance of the claim that our society would have been inundated by paper work without computers. Without computers, people would work a little harder and get a little less for their work. However, when home terminals become available, social changes of the magnitude of those produced by the telephone and automobile will occur. I have discussed them elsewhere, and I think they will be good - as were the changes produced by the automobile and the telephone. Tyranny comes from control of the police coupled with a tyrannical ideology; data banks will be a minor convenience. No dictatorship yet has been overthrown for lack of a data bank.

One's estimate of whether technology will work out well in the future is correlated with one's view of how it worked out in the past. I think it has worked out well - e.g. cars were not a mistake - and am optimistic about the future. I feel that much current ideology is a combination of older anti-scientific and anti-technological views with new developments in the political technology of instigating and manipulating fears and guilt feelings.

3. **What motivations will artificial intelligence have?**

   It will have what motivations we choose to give it. Those who finally create it should start by motivating it only to answer questions and should have the sense to ask for full pictures of the consequences of alternate actions rather than simply how to achieve a fixed goal, ignoring possible side-effects. Giving it human motivational structure with its shifting goals sensitive to physical state would require a deliberate effort beyond that required to make it behave intelligently.

4. **Will artificial intelligence be good or bad?**

   Here we are talking about machines with the same range of intellectual abilities as are posessed by humans. However, the science fiction vision of robots with almost precisely the ability of a human is quite unlikely, because the next generation of computers or even hooking computers together would produce an intelligence that might be qualitatively like that of a human, but thousands of times faster. What would it be like to be able to put a hundred years thought into every decision? I think it is impossible to say whether qualitatively better answers would be obtained; we will have to try it and see.

   The achievement of above-human-level artificial intelligence will open to humanity an incredible variety of options. We cannot now fully envisage what these options will be, but it seems apparent that one of the first uses of high-level artificial intelligence will be to determine the consequences of alternate policies governing its use. I think the most likely variant is that man will use artificial intelligence to transform himself, but once its properties and the conequences of its use are known, we may decide not to use it. Science would then be a sport like mountain climbing; the point would be to discover the facts about the world using some stylized limited means. I wouldn't like that, but once man is confronted by the actuality of full AI, they may find our opinion as relevant to them as we would find the opinion of *Pithecanthropus* about whether subsequent evolution took the right course.

5. **What shouldn't computers be programmed to do.**

   Obviously one shouldn't program computers to do things that shouldn't be done. Moreover, we shouldn't use programs to mislead ourselves or other people. Apart from that, I find none of Weizenbaum's examples convincing. However, I doubt the advisability of making robots with human-like motivational and emotional structures that might have rights and duties independently of humans. Moreover, I think it might be dangerous to make a machine that evolved intelligence by responding to a program of rewards and punishments unless its trainers understand the intellectual and motivational structure being evolved.

   All these questions merit and have received more extensive discussion, but I think the only rational policy now is to expect the people confronted by the problem to understand their best interests better than we now can. Even if full AI were to arrive next year, this would be right. Correct decisions will require an intense effort that cannot be mobilized to consider an eventuality that is still remote. Imagine asking the presidential candidates to debate on TV what each of them would do about each of the forms that full AI might take.

References:

McCulloch,W.S.(1956) "Toward some circuitry of ethical robots or an observational science of the genesis of social evaluation in the mind-like behavior of artifacts." *Acta Biotheoretica*,**XI**,parts 3/4, 147-156

Weizenbaum, Joseph (1965) "ELIZA—a computer program for the study of natural language communication between man and machine", *Communications of the Association for Computing Machinery*,**9**, No. 1, 36–45.

John McCarthy
Computer Science Department
Stanford, California 94305