Prospects for the Cyberiad: Certain Limits on Human Self-Knowledge in the Cybernetic Age

JOHN BARRESI

If the human brain were so simple that we could understand it, we would be so simple that we couldn’t.

—Emerson Pugh

1 Introduction

In his delightful book, The Cyberiad, Stanislaw Lem (1976) depicts a distant future where humanity no longer exists; however, man’s impact can be seen in the ever evolving, complex societies of cybernetic men which span the galaxy. The continued evolution of these inorganic machines is guaranteed by the existence of constructors: machines whose primary function is to produce new machines according to the specification of their clients — other cybers — or for pure pleasure in their competitive activities with each other. Some of these new men fail to pass the test of time due to their singlemindedness in achieving particular goals — which often lead to hilarious consequences. But others do last, and result in the development of ever more complex machines and relationships among machines. At several points the cybers discuss the myth that men were once made from flesh and blood. Some cannot believe this story — for they can hardly imagine the point of such a creature, let alone believe that it might have been the source of the cybernetic race.

With the dawn of cybernetics and the computer explosion, we seem to be rushing headlong into the future, turning science fiction into fact. The narcotic thrill of enthusiasm and optimism in the cognitive sciences now runs so deep that many believe a cybernetic machine comparable to man can be constructed within a few centuries and certainly within a millennium. But is this optimism warranted? Moreover, is the Cyberiad itself a real possibility?
If mechanism is right, why not? Aren't we just machines ourselves? Isn't organic chemistry just the inorganic chemistry of carbon compounds? Isn't DNA just an aperiodic crystal, as Schrödinger suggested in his What is Life? (1948)? Isn't an organism just a "gene machine" (Dawkins, The Selfish Gene, 1976)? Isn't the body nothing but a cybernetic machine, the brain nothing but a sophisticated computer, and the mind nothing but a program?

If that's the case then we can replace talk of life and mind, consciousness and free-will with their formal equivalents. We ought to be able to create mechanical analogues of processes and replace living, conscious machines by non-living, non-conscious substitutes in the scheme of nature. It might take some time, but it ought to be possible to initiate the Cyberiad. Moreover, if the current wave of scientific optimism is warranted, man might just do it whether he wants to or not.

However, is it possible? Is it possible to create formal analogues to life and consciousness which are fully adequate as functional equivalents, yet neither alive nor conscious? Some would argue that any formal equivalent would, by definition, be alive and conscious since it performs the very activities by which life and consciousness as concepts are necessarily defined. But this need not follow. For instance, it may be the case that there is a large variety of alien life forms in the universe which form a natural kind with life on this planet. The attribute which we share with these forms may be an inherent property of matter of certain kinds in particular organizations. If this is the case, it might be possible to mimic the functional structure and behaviour of these forms in yet another medium without producing this emergent property called life. Nagel (1979) has made a similar claim for consciousness. He suggests that there may be some subtle attribute of matter which in some combinations produces consciousness as an apparently emergent property — a sort of "mental chemistry" of point of view. If life and consciousness are real, inherent properties of matter, it may be that non-living, non-conscious analogues of a self-conscious intelligence like man, such as depicted in The Cyberiad, may be an impossibility. Moreover, even if possible, their principled construction may require an intelligence far beyond man's capabilities.

Nonetheless, I think it is an empirical question and one well worth asking. Can man initiate the Cyberiad? The question is interesting not because it is a worthwhile endeavour. Rather the interest lies in its use as a probe to investigate the possible limits of human self-knowledge, and to help define what it means to create an artificial intelligence equivalent to man. In the present paper I wish to propose a new test for artificial intelligence based on The Cyberiad. Let us call it the Cyberiad test. In the next two sections I will describe the test, compare it to the Turing test (Turing, 1950), and suggest how it might be used to evaluate the question: Can man initiate the Cyberiad based on knowledge of himself? In subsequent sections I will discuss a variety of difficulties that make it seem unlikely that we will ever pass this test.

2 The Turing Test and the Cyberiad Test

In the Turing test (Turing, 1950) a computer program is typically used which simulates human intelligence. The test involves fooling a person into believing that he is communicating with another real person when he submits verbal questions via written or typed messages to the computer program and receives answers from it. If the person cannot distinguish between answers given by another human being and those given by the program, then the program passes the test. The intelligence of the program is presumably indicated by its ability to fool the person in a situation where the perceptual properties of the machine on which it runs cannot bias the outcome. The test is clever, has been very useful in the past, and will continue to be useful in the future. However, it has several important limitations as a test for intelligence which have caused problems that can be avoided through the use of a more stringent criterion of intelligence, such as embodied in the Cyberiad test. So, before describing these problems, let us consider this test.

The Cyberiad test was constructed in an attempt to provide a behavioural criterion of human-like intelligence. It is a test of general intelligence, rather than being restricted to a particular domain — such as verbal communication. It assumes that human-like intelligence can only be clearly demonstrated by constructing machines capable of replacing human beings in a social setting. For instance, at a bare minimum a criterion that a cybernetic man would match my behavioural intelligence is if he were to replace me in society with no noticeable difference in our performance of comparable role activity. Of course some functions of society would necessarily change. In particular reproduction would have to be replaced by production of replacements. And the Cyberiad test itself depends for its justification on this one major difference between a human society and a human-like or cybernetic society.

My cybernetic substitute must also be able to function successfully in a social network of linguistic beings — a network of beings who communicate in a language which is ever evolving and hopefully increasing its hold in reality. It is clearly the case that human culture carries knowledge from one generation to the next, and seems constantly to expand the total amount of knowledge so carried. Much of this knowledge is carried not so much in the minds of individual men but in the language through which they communicate. It is a shared and ever evolving form of knowledge not
owned by the particular, but by the collective. And only if one is a participant in this collective, can one come close to fully appreciating the meanings of the transmission of signs or symbols among the individuals. While our current understanding of this process of linguistic evolution is limited, unless our cybernetic substitutes can engage as successfully in this process as ourselves, it hardly seems reasonable to suppose one has produced anything like a human-like intelligence.

I stated earlier that at a bare minimum my cybernetic substitute should show no noticeable difference in performance to myself. However, this is in itself an inadequate criterion of my intelligence. In comparison with the proposed Cyberiad test, it fails in one crucial respect. Indeed, it is nothing more than an expanded version of the Turing test, where now the program is hooked up to moving machine, and given a more general test of its intelligence. While this is some improvement over the original Turing procedure, it does not represent any dramatic shift in our criteria of intelligence. We still depend on a human judge as to whether the cybernetic man adequately simulates my intelligence. The Cyberiad test replaces this potentially limited perspective of our intelligence by letting mother nature, who constructed us, herself decide how well we have constructed our possible replacements.

The Cyberiad test depends critically on the following consideration: If I, as a cyberneticist, can create a cyber to replace me in society, then the cyber is necessarily intelligent enough to replace himself as well. Strictly speaking, I myself need not be able to build my replacement. What is critical is that my society or culture as a whole has acquired the knowledge necessary to build my replacement. I, as a cyberneticist, am only one member of this collective which shares the knowledge among its participants — no one individual being fully responsible for all the knowledge. What is critical to the Cyberiad test is that I and every other single member of this collective can be replaced by a human-like cybernetic substitute based on collective knowledge. If such is the case, then, in principle, every one of the human beings which originally formed the collective can be replaced one by one by cybernetic substitutes. Hence, once humans have achieved the ability to construct cybernetic men that are their full equivalent in intelligence, these cybernetic men could then replace man and should be able to carry on recursively the construction of their own replacements without further human assistance or even existence. Moreover, just this general social activity of the maintenance and replacement of cybers seems, unless greatly restricted in its conceptualization, an adequate criterion of human-like intelligence. Hence, the Cyberiad test imagines such a society as the goal of our effort to create an artificial intelligence which is our equivalent.

Before going further, let us review the logic which grounds this test. My goal has been to develop an assay of a human-like artificial intelligence, which is fully adequate to the job. Now, I am supposing that the purpose of this test is to provide an algorithm for evaluating any human-like intelligence which we have produced. (Or any concrete proposal on how we might produce one.) Hence my first assumption is that human beings have produced the human-like cybernetic substitutes. My second assumption is that these cybernetic substitutes are, indeed, capable of replacing us individually in our human society. Now, given these two assumptions it necessarily follows that every individual person can be replaced by his cybernetic substitute. Moreover, since the intellectual competence of the substitute is hypothesized to be the same as the original, then any activity of the original (except those related to reproduction) can be accomplished by his substitute. This includes all the procedures necessary for the construction of each and all the cybernetic men themselves. This being the case, then all the human beings can be individually replaced and the resultant Cyberiad ought to be able to continue recursively constructing and replacing its members in a procedure homologous to human reproduction. Since this constructive activity is a sufficiently complex social task to guarantee that the general intelligence of the cybers is being used, I suggest that if such a society can survive the likely lifetime of our species, then it’s members are roughly equivalent to us in intelligence.

Although the Cyberiad test might be used to evaluate real cyber’s abilities, this seems impractical. It seems more useful to imagine the test as passed if one can provide a concrete proposal on how one might successfully institute a human-like Cyberiad — one where humans no longer exist but are replaced by a society of artificial men able to continue a socio-cultural evolution of their own without disintegration over an extended period, say of several million years (the rough life-time of our species). 4 I envision the test being used in much the same fashion as Block (1981) has recently used the Turing test — a concrete model of how one might pass the test, with a defence against various objections of incompleteness. To be complete a proposal must discuss the types of knowledge required to make the creation of the cybernetic man a real possibility and how this knowledge is to be acquired and used in his construction. It must depict in detail how one might initiate the Cyberiad and how it is to evolve on its own without further human intervention.

Now let us compare the Cyberiad test with the Turing test. A main limitation of the original Turing test is that the only behaviour exhibited is verbal and the only stimuli received are likewise verbal. This has created many problems. For instance, Block (1981) has recently developed a very unintelligent algorithm which passes the Turing test. The conceptual procedure involves having a massive file of memory tapes where each sentence of English up to 20 words or so is stored — with an appropriate
response. Once the first question is answered by accessing its question, then a new series of tapes conditionalized on the first question are used. And this goes on for the short period of the test (Turing suggested a 5-minute test). Block uses this algorithm to argue that “behaviourism” must be false and “psychologism” true, since this very stupid machine can pass virtually any behavioural intelligence test. Although Block can make this argument stick — at least under a limited range of objections — it relies entirely on the verbal repertoire of the Turing test. Certainly no behaviourist worth his salt would accept such a Turing equivalence test as a behavioural measure of intelligence.

The Cyberiad test does not fall prey to Block’s algorithm. For one thing, Block’s algorithm is for a finite test and the Cyberiad test is virtually unlimited. Moreover, I think more critically, Block’s algorithm cannot deal with non-linguistic information nor with learning. Suppose we place the cybernetic man in a situation where he views some visual event and has to explain what is going on. Such a mapping of acquired information which is non-verbal to a verbal response, totally annihilates Block’s algorithm. And it’s not at all clear how he could improve it while keeping the program stupid. One possible solution is to provide the machine with a camera to process inputs. Assuming a finite list of propositional structures could be recovered from this input, something analogous to Block’s algorithm might be generated. (Block does discuss briefly this possibility.) But if the list of propositional structures is finite how can cultural and linguistic evolution occur? Or, perhaps, all learning is merely some combination of this ideal set of original propositions? To my knowledge neither Block nor anyone else since Wittgenstein’s Tractatus has proposed a method for identifying such an a priori set of propositions from which to generate codes for all possible forms of knowledge.

This leads to a second problem with the Turing test. Whether inputs are verbal or not, most uses of the Turing test presuppose that inputs can be processed based on their forms or syntactic shapes. In Fodor’s important paper “Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology” (1980), he argues for a computational theory of mind as the basic program for cognitive psychology where he supposed that mental events involve computational processing of structures based only on their forms. In so doing he relegates semantics, e.g., reference and truth, to some domain other than cognitive psychology. It seems to me that Fodor’s program is very much the state of the art in artificial intelligence, and that the reason he can brush aside what he calls a naturalistic psychology (one with truth and reference) is because the Turing test only presents a solipsistic medium in which to determine the structure of the world or what he calls “Dasein”. In contrast to this Turing test of reference (cf. Putnam, 1981), the Cyberiad test requires a full naturalistic semantics. Responding to the world’s ontological structure based on the syntactic structure of transduced inputs — without a full-blown world model of situational context — is hardly likely to succeed, particularly when the cyber’s response also occurs within that whole-world context.

The third problem with the Turing test is that it measures intelligence in terms of the program’s ability to trick a person. But natural intelligence in its full richness involves flexible adaptation in the real world — where the “person” who must be convinced is mother nature herself, the ultimate arbiter of intelligence. And mother nature has an entirely different set of criteria by which to judge intelligence. For one thing, she doesn’t ask one question at a time and kindly wait a reply. Rather she bombards the organism continuously with hundreds or even thousands of questions in parallel. And most of the intelligence of the organism is expressed in intuitive and even reflexive responses which require no cogitation at all. While AI people are impressed by a program which plays chess, mother nature would be much more impressed by a cyber who can play a good game of tennis. It seems to me that the Turing test strongly biases us to think of intelligence as the intellectual activities of some cognitive demon, when in fact the most difficult cognitive tasks we face daily are the simple coordination of perception to action, where we recognize where we are and know what we are about as we face the innumerable obstacles and opportunities of life. And these cognitive activities represent only a small fraction of the intelligent activities of the organism as a whole.

The original purpose of the Turing procedure was to provide an objective assessment of the intelligence of machines. The Cyberiad test has the same purpose. The major difference is on how “intelligence” is to be objectively defined and how it is to be assessed. In the Turing procedure humans decide what represents a meaningful test for intelligent behaviour and by comparing the machine to human performance they also judge how well the machine passes the test. In the Cyberiad test intelligent behaviour is defined as those behaviours necessary for the cybernetic society to survive. Presumably these will be behaviours similar to those exhibited by humans in society. And the judge of this test is not mother nature, the source and judge of our own intelligence. Hence while the Turing test defines and tests the intelligence from a human point-of-view, the Cyberiad test goes beyond this intermediate definition of intelligence to define it from the point-of-view of nature herself.
I expect the Cyberiad test to be used primarily as a Gedank experiment. A person who believes that he has a fully adequate set of procedures for passing the test provides a description of these procedures in a way to convince us that he could pass the test. Such a description has been given by Block (1981) in the case of the Turing test. The purpose of the description and his dealing with various objections is to convince readers that he could, indeed, pass the test if he followed through with his procedure. Of course, practically speaking, it might take several lifetimes and more, for instance, to set up Block’s procedure. But this is unimportant. What is important is that the story-teller be able to convince his audience that he has taken into account all foreseeable problems that might prevent him from succeeding in following his plan.

Now, in using the Cyberiad test we must think ahead much further than for the Turing test. But since we don’t have to actually be there at the end of the several million years, all that really matters is that we can imagine ourselves there and to evaluate what we think would have happened by that time, given the proposed procedure for generating and maintaining the Cyberiad for this length of time. Likewise with the knowledge that the author proposes that cyberneticists will have by the time of the initiation of the Cyberiad. We have to decide, individually, whether the procedure for acquiring the knowledge he wishes to provide an account of is adequate. Moreover, it isn’t essential that he provide an account of all necessary knowledge as the test can be restricted to special problems that seem especially interesting or difficult. For instance one proposal might focus on the problem of how to guarantee that a Turing-equivalent cyber, ergo one who passes the Turing test, is also a Cyberiad-equivalent cyber, ergo one who, if replicated perfectly numerous times, would still be a member of an active Cyberiad several million years later. Another proposal might focus on the problem of education and cultural evolution in the Cyberiad. A third might focus an error correction over the length of Cyberiad of interest (the reliability problem). Finally, one might try to account for all of these problems and others at once.

In all of these cases I believe it is important to have on hand a population of humans to compare with the cybers. For this group I suggest that we imagine a parallel society of humans that do not reproduce in the usual recombinatory fashion, but that produce offspring by cloning. Since this procedure is comparable to that used by the separate cyber society, we can decide fairly the relative survivability of humans and cybers. To the extent that we can expect the human society to develop, change, and survive we should require the same result for the Cyberiad.

However, there are some difficulties associated with this last suggestion. For one thing, there are the “doomsday” predictions suggested in Note 5, above. It seems absurd to concern ourselves with a several million year life for the Cyberiad when we are likely to annihilate ourselves long before then. While self-annihilation may indeed occur, I do not see this as a problem for the Cyberiad test itself. For, if man is predicted to annihilate himself by some individual, then he should expect the Cyberiad also to annihilate itself. The more interesting question here is why anyone might believe that we could ever have the capacity to initiate a Cyberiad that can mimic our own future self-destruction, yet be unable to prevent it from happening. For myself, I have greater faith in humanity’s ability to attain sufficient self-control to prevent such termination of our species, than in our ability to duplicate it mechanically by initiating the Cyberiad.

A second problem associated with my suggestion for parallel societies is that I do not see how we can guarantee that the human society will maintain its practice of cloning. While it’s easy enough for me to make this suggestion, I have little faith that humanity would continue to pursue this policy for the several million years of the test period. During this time humanity would be expected to have fully populated this planet and to have moved on to other greener pastures in our galaxy. They will have their own problems to worry about and have little concern for our needs for a parallel society. And cloning, after all, may turn out to be a rather boring policy after a while. So I suspect that the humans could not be kept back from engaging in recombinatory genetics in the end. I really don’t have a realistic solution for this problem other than to accept this fact and admit a non-parallelism here. Or alternatively, the Cyberiad test users might imagine a human subpopulation that is particularly interested in maintaining the cloning policy for one reason or another and can be imagined to last the life of the species.

At any rate there can never be a perfect parallelism or even isolation of the two species in the end, were the Cyberiad to be initiated. So in effect we must put some constraints on the realistic use of our imaginations in considering the test, and the two species. Since we have to do so, we might as well imagine them as existing in two isolated ersatz possible worlds to prevent their interaction.

Furthermore, I think it useful that we imagine that no “accidents of nature” occur that would tend to aid or inhibit the existence of one of the two species over the other. What we are interested in is the alternative evolution of these species behaviourally, not in terms of effects due to their physical differences. So the impact of events that would differentially affect the two species based on their physical rather than than behavioural differences should be ignored.
Over the next six sections of this paper I wish to consider a variety of difficulties that we face in trying to pass the Cyberiad Test. Some of these problems seem to me to more certainly prevent the success of any attempt to initiate the Cyberiad than do others, but all of the sections are aimed at pointing out where important difficulties lie. It is my hope that each of these approaches to the difficulties as well as novel problems will become better clarified through future use of the test by other authors.

Let us now consider some of the empirical difficulties in initiating the Cyberiad. What we have to imagine is that our scientific knowledge of man has advanced far enough that we can create mechanical duplicates of each and every one of ourselves. As pointed out earlier, it isn’t necessary that each man be able to duplicate himself, but that society at large has the distributed knowledge to achieve this task — that each member of the society can, in principle, be functionally duplicated and replaced by a cyber-equivalent. How is this going to be possible?

First, let us consider the case where we believe that we can create physical duplicates of ourselves. By this I mean that we have, or believe we have, the physical knowledge of how our bodies and brains are constructed, and have now created functional equivalents of several humans, that to all appearance seem to behave like their originals — ergo they are Turing-equivalent cybers. The kind of knowledge that went into their construction was based entirely on our physical understanding of the brain and body and what we have done is to mimic these physical activities using other materials that we have no reason to suppose are “life” or “mind” endowing as essential properties (cf. Searle, 1980). Rather, what we have done is create duplicates entirely based on materialistic or physicalistic considerations.

Now, the first point I want to make in this case is that one has to accept on faith that such a physicalistic parallelism of functions can be achieved, between the kind of matter and its combination that produces “life” and “mind” as we know it, and an alternative that by our traditional essentialistic definitions are mere physical mechanism without “life” or “mind”. A faith in physicalism is not necessarily a faith that such a program will be possible. Rather, it is a faith that we will ultimately understand fully what makes it so that such qualities of matter as life and mind are possible. But let us grant at least the possibility in the present case that such physicalistic functional equivalents are possible, so that we can evaluate, even given these presuppositions, whether our Turing-doubles are Cyberiad-doubles.

The essential problem that I see in this case is the problem of entropy, and of reliable duplication over the period of the Cyberiad’s existence. What we have are physical objects created in such a way that they are composed of parts that are functional equivalents of parts of our bodies and brains. We are not to evaluate their Cyberiad-equivalence based on the global behaviour of these men, but rather on the operation of the parts out of which these men are made. As far as we can tell their global behaviour passes the Turing test — they appear to be our equivalents globally. But we have no theory of their psychology per se, so we are not evaluating their behaviour vis a vis such a theory. Rather what we have is a detailed specification of the physical functions of the human body and brain and the ability to duplicate such functions using other material, of a non-living sort, assuming that it is possible to mimic such processes. But this means that our criteria of duplication cannot be based on the mereological sum of the activities of each of our functional elements, but based only on those elements themselves. Hence our task is to guarantee that these elements are functional equivalents to the significant elements whose mereological sum produces ourselves, and to ensure the reliable operation of these elements. Assuming that we have the right physical theories, we presumably can guarantee that the elements that we use as functional equivalents perform the same functions as the original elements in our bodies and brains. But it’s not clear that we can guarantee that the idealized functions that we are matching are the significant functions whose mereological sum we are. Despite the evidence of first approximation through the Turing test, it may be that our models have not identified the true elements, but are nevertheless succeeding because of an approximate match. If this is the case and if our error correction procedures are aimed at maintaining these pseudo-functions within the limits set by their parallels in the human organism, then we have no guarantees that error correction of the true functions are equally equivalent.

The point I’m trying to make here was discussed following von Neumann’s (1948) Hixon lecture on Automata theory. It is the problem of reliability of operation and self-correction of errors in automata. von Neumann suggested that as long as one had a general objective criterion of a function of the nervous system, it was possible to control error and maintain reliability of the simulated functions, to within any level of accuracy that one wished to impose (see von Neumann (1948) pgs. 320–324). However, what is critical to von Neumann’s remarks and to the questioners’ doubts, is our ability to provide an objective criterion of what the functional description of the behaviour of the machine is supposed to be. In the present case I question whether we can guarantee that we have the right functional description of the whole organism when we focus only on the functional descriptions of what appear to us to be its significant parts. It may be that we have chosen the wrong elements, and although we
can constrain the limits of error of these elements, the cumulative errors over time relative to the true functional description will lead to entropic dissolution of the Cyberiad. I am uncertain of how fatal this objection is, but it seems at the very least to demand that our physicalistic, functional description be the right one, and it’s not clear how we can be sure of that based only on the limited testing of the operation of their mereological sums that are acquired via the Turing-test. Indeed, by focusing on the parts rather than on the whole organism, it seems that any efforts at setting criteria for error correction will be more prone to cumulative effects of theoretical miscalculations, than if we had directly aimed for some global criterion, such as we will next consider.

An alternative, or perhaps supplementary, approach to our empirical problem is to try to develop an adequate psychology of man, so as to define a behavioural criterion of equivalence where the organism’s behaviour is viewed globally rather than in terms of particular motions or actions, of its parts. What would be required here is a functional description of the whole man with particular focus on his global behaviour. There is at least some hope that some such global criterion might be generated through biological and psychological theorizing. And perhaps criteria from such sciences, partially independent of the physicalistic approach, might converge on criteria by which to evaluate whether we had, indeed, created cyber-doubles that could pass the Cyberiad test once they had passed the Turing-equivalence test. But how are we to make such guarantees? From biology’s point of view we need to specify exactly what the functional meaning of each body function and global behaviour of the organism is for its adaptation. Moreover, we need to do this separately for each human that we wish to duplicate. If we had such a theory, we could certainly apply appropriate von Neumann error correction procedures to guarantee that our duplicates stayed within tolerance over the period of the Cyberiad test, hence pass the test. But first we need the theoretical description of our functional structures from which to base this criterion. How are we to develop such a theory prior to our initiation of the Cyberiad?

It seems to me that such a biological functional theory requires, at the very least, that we are able to decode the adaptive significance of the string of DNA that defines each person as a potential organism. If we had such a functional theory of DNA we would have, in effect, reconstructed the evolutionary history of selective advantages entailed by each code, and so know how each code aids in the construction of the organism as well as determining the functional meaning of each of the organism’s behaviours. Is there any hope that such a functional theory can be achieved prior to initiating the Cyberiad? I think not. The basic problem here is that the string of DNA cannot be viewed as a series of independently coded functions. Natural selection treats organisms or strings of DNA as wholes. Success or failure depends on reproductive success of the organism, not of independent genes or elements of DNA. As a result, the biological functional description of a given, non-identical individual cannot be determined based on knowledge of each of the separately considered genes. It’s their interaction that determines the functional description of the whole. So at best it seems that only duplicates of each individual can be used to achieve a final description of his functional structure. But if this is the case, then we have only the individual and his identical twins to work with in determining the theory of his functional description. Can our biological science progress to the point where we can be sure based a Turing-equivalent passage of our biological functional theory that we had also ensured Cyberiad-equivalent passage? I see no reason to imagine that such would ever be the case no matter how many past individuals our biological theorizing had seemed to adequately simulate based on their single lives. For the error-correcting element in our theorizing is constrained to one lifetime, and we have not even tested it beyond one lifetime (or perhaps several with twins or cloning). So our behavioural criterion hardly warrants prediction over the multiple generations of the Cyberiad.

Does it help any to approach the individual as a psychological rather than, or supplementary to, a biological one. Instead of looking at final ends in terms of adaptive significance, we might approach him in terms of his proximate purposes and goals, in terms of the operations of his mind. It seems, at least, that such a psychology of individuals is a less ambitious goal for our ultimate science than a biological functional theory. While this certainly seems to be an easier task, our empirical problem is essentially the same. We can induce a psychology for people in general and for a particular individual as well, but how can we be sure that it provides a sufficiently true functional description and criterion of behaviour so that a Turing-equivalent passage of our theory guarantees a Cyberiad-equivalent passage? So long as our psychological theory is based on inductive generalizations, I don’t see how this would be possible. But perhaps we could develop a deductive theory of human psychology that applied to all humans and could be tailored to the individual case? Such at least seems to be the aim of cognitive science, and perhaps it will succeed. We will consider this possibility in the next section.

To summarize the present section: The basic empirical problem is to provide an objective definition of behaviour by which one can get beyond Turing-equivalency to Cyberiad-equivalency. One possibility for such a definition is to find such a criterion based on the mereological sum of criteria that are physicalistically based. While it does seem feasible to develop deductive physical science to the point that we can provide objective criteria for error correction of component parts of the human
organism, the difficulty here seems to be whether these are the true functional parts, relevant to global behavioural descriptions. If we tackle the global descriptions themselves, we can try a biological or psychological approach. Based on what we know already of the biological specification of behaviour, it seems unlikely that we can ever define the functional significance of a human organism’s behaviour in all situations that he might conceivably face over the multiple generations of the Cyberiad’s existence to pass the test. As far as current biological science can determine, it is the process of natural selection operating on the genetic material contained in DNA which carries the functional descriptive basis for a human organism. Since reproductive success is based on the holistic sum of the DNA codes, the adaptive meanings of these codes must be evaluated in context. Hence, each individual (and his twins) must be treated as unique. Thus, we have only a finitary base upon which to develop laws of behaviour for the individual, and hence only inductive generalizations of the functional significance of his behaviour seem forthcoming. Any deductive model sufficient to set behavioural criteria for our model of a man to constrain error over the lifetime of the Cyberiad cannot be obtained over the few lifetimes that our pre-Cyberiad scientists have to work with. An alternative holistic approach that might prove more successful is a psychological one. In this approach behaviour is treated within an intentional/psychological framework and our science tries to develop generalizations relevant to mind in general, and in the particular. But as far as such an approach is inductive, it seems that the limited temporal resources by which to generate objective behavioural laws of the individual negates the possibility that any Turing-equivalent double could be guaranteed to be a Cyberiad-equivalent double, since the latter must face situations over many generations on which our model has been insufficiently tested to ensure passage, within controlled limits of error.

5 Can man initiate the Cyberiad based on formal knowledge of himself?

If mind is nothing other than a program, the answer to this question appears to be “no”. I am not denying that cybernetic machines capable of replacing human beings in the scheme of things could, in principle, be constructed. Nor even that we might be ultimately responsible for such a development. What I am denying is that we could accomplish this feat through the development of a formal theory of human nature — in particular a formal theory of human psychology. And I think the problem arises from the fact that we are simply not in a viable position to accomplish this task.

Let us suppose that mind is a program and that the role of psychological theory is to develop a formal theory of these programs: a theory which generates all and only those programs that are human minds. Furthermore, in order to keep the focus of our discussion on psychology, let us assume that the construction of cybers to replace men depends only on an adequate psychology, since for any given “mental” program a cyber could easily be constructed out of durable materials to perfectly instantiate the program.

But can we complete such a psychology? I think not. For if mind is a program, then we are asking a formal symbol processing system to fully represent its own structure, and this it cannot do if it is a consistent formal language sufficiently rich to generate the arithmetic of the natural numbers — which it must be. This follows from Gödel’s second incompleteness theorem, which must apply to minds if minds are programs, since all programs — as we currently conceive of programs — are reducible to Turing machines and Gödel’s theorem applies to these machines.

The argument goes roughly as follows: One of the things we want such a formal psychology of mind to capture is an individual’s cognitive competences. So let us consider the arithmetic and logical competence of some person named Oscar (cf. Anderson, 1989). Insofar as his knowledge of arithmetic is true, it can be captured by the formal axioms of Robinson arithmetic, theory Q. Consider the extension of Q that identifies the formulas expressible in Q that Oscar is capable of proving. Now, construct a formal system whose theorems are the sentences derivable from the axioms of Robinson arithmetic together with the sentences that Oscar can prove. Call this system Q’. Under the assumption that Oscar’s arithmetic competence can be captured by some program and is hence Turing-reducible, then the sentences that he can prove are recursively enumerable, and the extended system Q’ is a consistent formal system capable of expressing arithmetic truths. Since this is the case, Gödel’s theorems apply to it and it can be shown that the system’s consistency cannot be demonstrated within the system itself, despite the fact that it is clearly consistent. The upshot is that if Oscar himself were faced with the metamathematical proof we have described, he could not assimilate it as a proof.

Now consider the possibility that I am Oscar. Through empirical procedures I believe that I have constructed a formal theory that represents my arithmetical competence; Q’ is the Turing Machine that captures this theory. But is this possible? Could I recognize Q’ as the theory of my own competence? In particular could I follow the proof of its consistency? On the one hand it seems that I just did — assuming that I have Q’ in hand. But on the other hand Gödel’s second theorem proves that I cannot. Hence we have a contradiction. The conclusion that follows
is that if my arithmetic competence is a program that can be simulated by a Turing machine, then I cannot discover that machine, or at least, recognize its program as identical to my own.\(^5\)

This argument could be equally applied to mind-in-general as well as in the particular, so it represents a general limitation on our ability to formalize our intuitive mathematical abilities. The lesson to be learned here is not that our mathematical competence is unformalizable per se, but that at any given stage in the development of formal self-knowledge of our arithmetic skills, there will always remain intuitive, creative skills that have yet to be formalized, hence the development of formal self-knowledge is a never ending process open to humanity.

In the previous section it was argued that we could never develop an adequate inductive theory of psychology by which to initiate the Cyberiad, and we have just argued that if mind is a program then a formal deductive theory of mind is also unavailable to us. However, it might be conceded that minds are not programs — hence Gödel’s theorem may not directly apply. Nevertheless, even without Gödel’s Theorem, it seems inevitable that any rich language which attempts to represent fully its own truth (Tarski, 1956; Löfgren, 1979) or its own knowledge (Anderson, 1983) produces inconsistencies, and that this will apply to any language of mind.

6 Truth, language, and thought

If the previous analysis is correct then the acquisition of self-knowledge for the human race is a never ending process. No matter how much knowledge of ourselves that we formalize in some theoretical scheme, it will always be incomplete; and indeed, serve merely as a stepping stone to new creative acts and to even broader self-knowledge. In the present section I want to look at this process more closely and to speculate on how it works. To do so I will find it necessary to deal briefly with a wide range of issues of current interest in the philosophy and psychology of mind, without providing a definitive, and perhaps even clear, resolution to these topics. Nevertheless, it is my hope that the general view that I present is sufficiently coherent as a whole to warrant this discussion.

As I see it, the general problem of “Truth” is to explain how it can happen that humanity, through the process of cultural evolution of language, increases its grip on reality over time. Viewed in this way, Truth involves a mirroring between language and reality that evolves over time, such that the terms of a language refer to more well defined entities over time, and such that culturally acceptable assertions involving such terms not only gain pragmatic effectiveness for action, but also become increasingly culture-free expressions of human phenomenological reality.

This general position is of the sort described as internal realism by Putnam in his book, *Reason, Truth, and History* (1981).

By associating Truth with language in this way rather than with an individual’s thoughts or particular theories, I wish to emphasize the cumulative and social nature of humanity’s ability to successfully use this term to describe a property of their experience. What makes truth a real property on this account is that, barring events such as the destruction of the Alexandria Library, humanity is capable of cumulatively increasing the mirroring property of its linguistic representation of reality, through social communication and the evolution of language. And the capacity for this activity, the generation and preservation of “Truth” in this sense, is an activity we must provide an adequate account of, if we are ever to succeed in creating artificial men anywhere near our capacity for evolving self-knowledge.

Given this approach to Truth, we can now ask more concretely how it comes about — how does it happen that a socially communicative organism, increasingly utilizes terms, and sentences, whose propositional content increasingly mirrors or represents reality? In Frege’s (1956) terminology: how can the “Thought” come to refer to the “True”? With Frege I also take the thought of a sentence to be in general an objective interindividual and ultimately cross-cultural property of the meaning of sentences. As a language becomes more perspicuous, its terms and sentences not only increase in effective reference to true states of affairs, but the objective thoughts that they carry become better articulated representations of that reality. But as Quine (1953) made clear, we must view this mirroring process of language holistically rather than fixing our attention on individual terms and sentences, no matter how basic they seem to be. Hence, it seems reasonable to couch our question in terms of a Tarski hierarchy of languages, where metalanguages recursively define lower order object languages in terms of themselves. Quine (1969) speaks of the “background language” in which any theoretical language’s terms and sentences get their references fixed as such a Tarski-like metalanguage. Let us suppose then, that human language as it evolves, and indeed human thought as it evolves, can be viewed as a series of metalanguages which fix the truth and meaning of earlier languages in terms of themselves. Our question now becomes: How do metalanguages grow out of each other so that a previous metalanguage becomes the next object language of a more embracing picture of reality?

If we only had language itself to deal with, then this would be a complete mystery — indeed, an impossibility. Consider any language you want and the reality it describes however imperfectly. If the resources of that language were only redefinitions of words of that language in terms of other words, no increase in fit between that language and extra-linguistic
reality could ever emerge. Indeed, I believe that the meaning of the Löwenheim-Skolem theorem as applied by Quine (1966, 1969), Putnam (1981) and others to show how arbitrary the fix between language and reality is, merely demonstrates that without extralinguistic resource in human phenomenological experience upon which to base metalinguistic growth, the truth preserving or mirroring property of a language is fixed. And, unless I’m deeply mistaken, this is going to be one of the most difficult problems to deal with in attempting to create cybernetic duplicates to ourselves. For we must give them the wherewithal to develop new metalanguages, out of a non-linguistic experiential base which parallels human prelinguistic phenomenological experience. Yet this is going to be difficult to do unless we can articulate linguistically what that base is in ourselves. But if we could do that, then something like Wittgenstein’s Tractatus program would have to be fully cashed out, and even for ourselves our metalinguistic capacity to articulate experience would have reached its upper bound, and no more metalanguages could be expected to emerge through the dialectic of prelinguistic or phenomenological and linguistic experience. However, as we saw in the previous section such an ultimate formalization of possible experience through language is an impossibility, prevented at the very least by our inability to represent in that language our full capacity to generate new arithmetic truths, and through these new discoveries to expand formal mathematical language by which we can generate even further truths.

Assuming then, that in our own case new languages emerge out of a dialectic of human non-linguistic and linguistic experience, I now wish to investigate this process to see how truth comes about. It will be useful to our enterprise to consider at this point two distinctions. The first is between “information” and “representation” (cf. Dretske, 1983; Sayre, 1986). Information is viewed here as a structural residue of causal interaction with other physical objects that is carried by some object and can be recovered from it given the appropriate framework through which to interpret this residue. An example of such an object is the human brain which carries structural residues of its causal interaction with objects and events that happened in the life-history of the person (as well as, in its basic organization, the life-history of organisms on this planet). As this example indicates, and as the example of DNA in section 4 also indicates, there may be much more information contained in an object than can be recovered from it. This is because information stored in these structural residues becomes useful information only relative to a framework capable of unpacking the information and converting it into knowledge. And this is where information differs from “representation”, for representation in some real “sense” identifies its own truth conditions, hence presents its own framework for evaluation. Representation is a semantic concept whereas information is a material concept. Representations can be wrong, they can misrepresent, whereas information can never be wrong — it can only be misread, or misinterpreted. The two concepts connect the notion that a representation is “interpreted” information. As a result of this, we can regard representations as “true” or “false” depending on whether or not the projected conditions of this interpretation are satisfied by the information referred to in the representation. Although it is probably not as simple as this, we can tentatively define “intentionality” as a property of organisms by which information comes to be interpreted so as to become representation. Organisms with this property have “minds” as well as “brains”. The “mental inexistence” of the object of intentionality comes about from the fact that representations are about something, that something being the states of affairs that satisfy the “interpretive” conditions of the representation.

The second distinction that I want to make use of is that between a semantical theory of truth and a reliability theory of truth. In recent years there has been an increasing interest in what Field (1974, 1978, 1984) has called the reliability theory of truth (see also Loar, 1981). The basic notion of this approach is that we need an empirical theory to explain what is so special about truth in the semantical sense, why is “true” a useful attribute to apply to certain sentences in a language. Field’s suggestion is that “true” is a term that facilitates the useful transmission of reliable information — where information here can be taken in the sense above — ego as a carrier or residue of causal interactions. Without going into details, we can see that this distinction parallels our previous one in taking the truth property of representations as a means of connecting representations viewed semantically to representations viewed as material carriers of information from other aspects of reality. A useful reliability theory would explain the empirical role that truth plays in this relationship between representation and information, and it is to this that I now turn.

Since truth as I defined it earlier is to be viewed as the property of sentences acceptable in a language, our task now is to investigate how this comes about empirically and to see how it connects to the material activities of the human organism and hence to the informational properties of these activities. If we consider the process of language learning, it seems clear that the task of the child is to learn to respond to information in a manner that makes his verbal and non-verbal behaviour explicable within a language. What he or she says and does is being shaped in this process of language acquisition to fit the “truth conditions” of the language. In effect, he is being introduced into the group “mind”, for the public language is the language of thought for the social group. This is why Frege suggests that “thought” is a property of mind, not of minds (Frege, 1956). It is an objective or interpersonal representation of reality.
that is grasped through private "ideas" of the individual. The individual, as he acquires language, learns to use it in governing his behaviour, through the feedback of other individuals who share this language. But it is not individuals, but the language itself, which carries the objective thoughts of that language, and only insofar as the language grasps reality, are we able to say that the culture itself is moving toward universal human truth. However, at each stage in its development, the public language itself serves as the criterion of truth in terms of publicly acceptable assertions. The growth of language comes about through the modification of the set of assertions that are taken to be acceptable. This in turn depends on the negotiation among individuals in their private interpretations of the language and of the reality to which it refers. As each individual participates in the dialectic between the phenomenological base of prelinguistic experience and linguistic processes, the language itself gets modified. What individual terms come to represent changes and vocabulary and acceptable assertions get added or deleted from the language. Through this process, information originates in each individual's processing of physical and social reality, and gets transmitted among the various participants of the language, gradually modifying the power of the language to represent this information, as the individuals attempt to improve their adaptive fit to physical and social reality.

This is not to say that languages necessarily move in the direction of a better fit with reality. As civilizations rise and decline there is not only a tendency for language to increase its fit to reality, but it can decrease as well. It is because the human mind cannot be treated in isolation from other minds that such wholesale changes in linguistic grasp of reality can occur. Nevertheless, insofar as there is continuity in languages, and past communicative acts that have been written down and retained are still interpretable so as to translate these past thoughts into present ones, then the hierarchical growth of metalinguistic knowledge is still possible, and not entirely lost through time. However, the success of any such endeavours requires that individual human beings be able to translate these past thoughts into present ones, which can only happen if the grasp of present thought is wider in scope than these past thoughts and can define them into current background language. But the ultimate background language is phenomenological experience itself and unless we can share phenomenologically the past experiences of men, we cannot fully grasp their thoughts. So in the end public language is not sufficient, for it is merely a tool by which we share in our common humanity.

The net result of all this is that unless we can create cybernetic men who can share with us the human point of view of reality that is independent of particular experiences, they will not ultimately grasp reality in the same way we do no matter how similar our language expressions and other

behaviour appear to be over a finite period of time. And as pointed out earlier, it is not clear that there is any hope for us to create artificial men who can share with us this perspective because the perspective is not itself formalizable in a language — indeed, it itself fixes the ultimate meanings of all languages. In the next section I will look more closely at this phenomenological base.

7 The problem of phenomenological experience

The problem which Kant (1781) was the first to face and the problem which very few others since him have even acknowledged, let alone faced, is the problem of how phenomenological experience is possible. To my mind there have however been at least two important philosophers other than Kant's followers that have contributed to dealing with this problem. They are Frege and Wittgenstein. In the present section I wish to briefly outline Kant's problem and his solution and to try to show how Frege and Wittgenstein have extended his work through their theories on thought and language.

Kant's Critique of Pure Reason attempts to deal with a number of issues, but certainly the most significant problem for him was to provide some account of how objective experience was possible at all. There are a number of ways to view this problem, but congruent with my general purpose, one can phrase it as the problem of how objective thoughts in Frege's sense are possible. How is it possible to develop senses or SINNE for language such that the language expressions given in context can express thoughts that have a truth value, where truth here is Fregean truth — objective in the sense that it can be had by any rational human being capable of grasping the thought, and objective in having a truth value relative to some aspect of the world as experienced by humanity? In other words the problem is how is it possible for there to be a public reality that is jointly grasped by minds yet also individually experienced by them. Kant's solution to this problem is that the form of our experience of reality is determined a priori — prior to experience, and that the reason we can acquire objective knowledge of it is that its form is the same for all of us, and determines our joint attempts to form judgments about it. Hence any thoughts we form based on experience are constructed out of concepts that mirror the apriori form of that experience. Without the apriori forms of experience there can be no common experience, nor objective thoughts of that experience. This is the view I take to be internal realism, which opposes itself to metaphysical realism. The net result of this point of view is that the dialectic between phenomenological experience and language described in the previous section is an adventure restricted to humanity,
or perhaps to some higher being capable of experiencing reality in a human way, but which almost certainly excludes any machine that we would ever be capable of constructing.

For consider the following: A human being's experience of reality is not merely the perception of something outside of himself. Rather it is an ongoing experience of himself as perceiving reality outside of himself — a reality which he can come to form objective knowledge about through his communicative interactions with his fellow man. These communicative acts allow him to experience the same reality from alternative points of view — yet, and this is critically important for seeing the contributions of Frege and Wittgenstein, these alternative points of view are not the individual psychological states had by these other individuals but rather "objective" thoughts had by them. As Frege and Wittgenstein try to make clear for us, the thoughts expressed through language, are not ideas, feelings, etc. of individual conscious experience, but objective in the sense of being defined via objective criteria in the common world of our experience. Insofar as we might have thoughts based on incomunicable criteria, these thoughts have no objective content that can be appreciated by any other individual, hence cannot enter into any objectification of reality. The net effect of all this is that one must be a human to enter into the human dialectical process by which phenomenological experience becomes objectified as public truth via language, and again I see no hope for us to create human-like substitutes who can enter into this same game — playing by the same rules that we do.

8 The material and the phenomenological

All of the foregoing leads to what is probably the largest problem that we face in trying to initiate the Cyberiad. This problem is how to convert our own phenomenological reality into a physical instantiation. It is one thing to create a machine which performs according to specifications that we can interpret within our own point of view. It is quite another to create a machine that shares with us that point of view. In the previous section I have suggested that this is impossible — but I viewed the problem phenomenologically. Now let us look at the same problem materialistically. The problem is to create machines that convert information into representations by the same apriori rules that we do. Such machines must convert material signals from the world-at-large into a model of that world which they can objectively share with each other and increasingly come to represent in their linguistic processes. The materialistic puzzle is how to guarantee that these machines share the same apriori model of reality — the same phenomenological base. The puzzle here is not how to get them to "perceive" the same things. Perception is a relatively simple affair so long as you isolate the percepts from each other. But what is required is not merely a unity of percepts but a unity of experience into which these percepts get fitted, where each machine finds that he has a space-time locus in this experience and he can verify differences between subjective and objective events relative to the same experiential model that he finds to obtain in his cybernetic brethren. How are we to guarantee that the information that each machine processes, ultimately produces the same reality for them as it does for us?

At the deepest level of our experience, we can distinguish between the subjective and the objective. We all are able to make sense of this distinction and can divide our experience into those which are private and those which are public. The private ones are those aspects of our experience that we take to be localized in our body. The public ones are those beyond the limits of this "inner" perspective. Now, our task is to get machines that make the same distinction. Moreover, and much more puzzling and critical, it must be the case that, unless a particular machine has gone crazy or is mentally incompetent, it must be able to verify when its own point of view of reality is in error relative to the other machines. It must be able to see in a primitive phenomenological sense when the error is due to its own subativity, relative to objective experience. To do this successfully it must be able to view itself "objectively", relative to a shared model of reality with others. This model furthermore, has its phenomenological base plus its shared linguistic interpretation, both of which differ from the momentary subjective perceptions of events as the machine interacts physically with reality as well as the social order of machines. But how can it do it?

One possibility is to suggest that since the same material source of information is involved for each of these machines there is no reason why they cannot come to experience it as coming from the same source. But this presupposes that they can fix some event or object in its space-time coordinate by checking with each other. If each machine has a reasonably good internal clock by which to organize its representations into an experience and also keeps track of where it is continuously and projects objects and events into space and time as we do, then presumably they might check with each other much as we ourselves do to verify the objective events of experience. Moreover, assuming that their experiential base is as similar to each others as ours appears to be, then they should be able to develop objective "thoughts" out of these experiences that are organized in their "unity of apperception" much as we do in our language acquisition and communicative processes. In other words, if they have an apriori phenomenological base of the same sort as ourselves which they share, then they can also come to objectify this apriori experience, much
as we do. But again it seems to me that while we might try to construct machines with an apriori phenomenological base out of which to create objective experiences, what we cannot do is to give them our base. For it is only relative to our base that we can even conceive of what is required and the rules for the apriori form of our experience will never be fully accessible to us — at least not in a language that is “materialistically” based and goes beyond that experience in order to fix it in a manner that we can duplicate in inorganic machines. At least, such a possibility seems inconceivable to me. But if others wish to believe it possible and to pursue it as a goal, I wish them good luck on their adventure.

9 The essential indexical

The problem that I wish to consider briefly in the present section is not an epistemological difficulty as such — though it no doubt has its epistemic side. Rather, it is an issue of ontology. After all is said and done in our attempts to understand how a human-like cyber might be possible, there is still one problem that — if real — seems insurmountable. This is the problem of “bearers” (cf. Frege, 1956). How is it possible for a machine that carries information to somehow become, or at least to mimic, a creature which bears representations? Given our previous analysis of “intentionality” as the property of organisms by which information comes to be interpreted so as to become representation, we can define a “mind” as a bearer of such representations. The representations it bears are then, perhaps, the medium of immediate existence of a mind. This medium is, in a sense, what it is to be a mind — to be a subject of experiences, whose polar opposite, the objects of experience, are the projected objects of the representations that we bear. If this is the case, then how is it possible to make machines that carry information, but do not “interpret” them so as to bear representations, act as we bearers do? Whereas the machine processes information and moves, we think, feel, believe, desire and act, based on our interpreted information. We make a choice while the machine calculates an outcome. Are the differences here real or are they just apparent? Does the ontological difference really matter, or is it perhaps merely a pseudo-ontology? Is intentionality a real phenomenon, or not? And even if it is real, does it make a difference?

To me it seems more or less a question of faith. Those of us who believe that intentionality is real, and that only bearers can have experiences through bearing representations, find it difficult to believe that non-intentional information processors could ever replace us in the scheme of nature. If human action depends on our subjecthood at all, it seems inconceivable that an object that lacks it could ever really “care” enough to do what is “right”. Unlike us, it has no relationship with other objects or beings like itself; it is essentially alone and doesn’t even know or care about it. So how can it behave as if it did?

Yet, my faith here seems, ever so easily, refutable. Haven’t we already built machines that mimic human behaviour so well that people communicating with them believe they are human? Even if there is some doubt here, we certainly have come close to this point, so how much further can it be until we can’t tell the difference? So maybe there is no difference to speak of between truly intentional and non-intentional beings — all that matters is the right kind of functional organization, independent of whether the bearing of representations is involved.

But this is where we began.

10 Conclusion

In the present paper I have proposed a new test for evaluating the intelligence of cybernetic men. The purpose of the test is to provide an adequate formal procedure for evaluating whether these machines truly mimic man. I have also discussed a variety of difficulties that seem to me to make it impossible for us to ever pass this test either actually or conceptually based on knowledge of ourselves. Some of the difficulties that I presented may be, in fact, only temporary setbacks that later we will be able to explain. Others might not be real difficulties, though they have appeared so to me. Finally, some of the difficulties might be real, and indeed prevent us from being able to purposely create cybernetic duplicates. Whether this is the case or not, and whether new and more important difficulties appear in the future, I hope that the test provides us with a relatively objective procedure for evaluating this possibility and for clarifying our thoughts about it.

It’s rising, flashing, mounting. In a jiffy it’ll be done. A great project seems crazy at the beginning, but the day will come when we’ll laugh at our luck. Sooner or later a thinking man will be able to make a thinking brain.

Goethe, Faust (pg. 119)

John Barresi
Dept. of Psychology
Dalhousie University
Halifax, NS B3J 4J1
Canada
Acknowledgement My initial thinking on the issues of this paper was first stimulated by Dick Smith's (1977) sabbatical colloquium. The present paper is based on a talk presented to the Philosophy Department at Dalhousie University on June 18, 1982. I would like to thank my colleagues in both Philosophy and Psychology for their helpful comments on that talk as well as on earlier drafts of this paper with special thanks to Jim Clark and Terry Tomkow. Research support was provided by NSERC.

NOTES

1 George E. Pugh. The Biological Origin of Human Values. (Quote pg. 154).

2 In answer to the question: "How would you find out if a computer or a machine was alive?" Pattee (1979) answers: "I believe it's useful as a strategy to take the widest possible point of view and say that if a computer could describe itself to the extent that it could also construct itself — in other words, if it had an internal machine shop which, following its own instructions, could construct another computer like itself — at least for the sake of argument, I would consider it to be a self-reproducing machine and alive. It could, in fact, under those conditions, evolve by mutation and natural selection, but I doubt if present society would support such activity." (pg. 91).

3 Searle (1980) makes a somewhat similar point in regarding brains as a special kind of machine — one which produces intentionality and other mental phenomena. "I offer no apriori proof that a system of integrated circuit chips couldn't have intentionality. That is, as I say repeatedly, an empirical question. What I do argue is that in order to produce intentionality the system would have to duplicate the causal powers of the brain and that simply instantiating a formal program would not be sufficient for that." Searle (1980, pg. 453).

4 I'm assuming that humans have existed as an identifiable species for roughly two million years and that we have two million more years before identifiable new species will evolve out of us. The fact that we might annihilate ourselves long before then is a real possibility that has not entered into this estimate.

5 A more rigorous version of the present Gödelian argument that I have presented can be found in Benacerrafl (1967) (I have also referred to Anderson, 1983, and Chihara, 1972). This argument has been criticized, but not refuted, by Chihara (1972) and Webb (1983). Chihara discusses his own improved version of the argument concludes, "... what the argument shows, at best, is that I can never discover 'my own program'". (p. 342) Webb says, "... all we get then is the rather unsurprising result that, if we are this or that specific universal turing machine, then we can't always be sure of which program we are using for certain tasks." (p. 345). Putnam (1984) has recently presented essentially the same argument based on our arithmetic competence, as well as, what appears to be a more generalized version which shows that we cannot develop a prescriptive model of our own cognitive competence. Anderson (1983) discussing the paradox of the knower considers related arguments.

Finally, it seems worthwhile to quote Gödel's 1951 opinion. He believed that "the two most interesting rigorously proved results about minds and machines are these: 1) The human mind is incapable of formulating (or mechanizing) all of its mathematical intuitions. I.e.: If it has succeeded in formulating some of these, this very fact yields new intuitive knowledge, e.g. the consistency of this formalism. This fact may be called the "incompleteness" of mathematics. On the other hand, on the basis of what has been proved so far, it remains possible that there may exist (and even be empirically discoverable) a theorem-proving machine which in fact is equivalent to mathematical intuition, but cannot be proved to be so, nor even be proved to yield only correct theorems of finitary number theory. 2) The second result is the following disjunction: either the human mind surpasses all machines (to be more precise: it can decide more number theoretical questions than any machine) or else there exist number theoretical questions undecidable for the human mind." (In Wang, 1974, pg. 324).

REFERENCES


