

THE “AI DEBATE”: ARGUMENTS AGAINST AI

Harold G. Slovic

Key Words:

AI, Artificial Intelligence, CS, Cognitive Simulation, Computer Science, Gödel's Proof, Phenomenology, Physical Symbol System Hypothesis

I. INTRODUCTION

Nowadays, almost everyone who reads a newspaper or watches the news on TV is familiar with the words “artificial intelligence” (AI). To many, however, these words may connote a “fait accompli”, something that has already been accomplished. However, there are some people who assert the contemporary non-existence or the impossibility of AI. As with other technological endeavors having the potential for profound impact on human life, there is little about AI that is free from controversy. Efforts towards its realization have been accompanied at every step with uncertainty, hope, frustration, failure, and active debate.

It is the purpose of this paper to examine some of the major phi-

losophical issues and arguments put forth by "the opposition" in the "AI Debate", in an attempt, thereby, to provide a basis for greater understanding of the intractability of the problems being encountered in the efforts being made to bring AI into existence. In doing so, this author plans to adopt John Casti's convenient classification of the types of arguments posed in opposition to the concept of AI:

It looks as if the main philosophical arguments opposing the concept of a computer's ever having a real thought come in three primary colors: *phenomenological* arguments based upon the belief that the totality of human understanding cannot be mechanized, *logical* arguments revolving around the limitations posed by Gödel's theorems, and *antibehavioristic* arguments founded upon the notion that behavioral observation alone is not enough to conclude the presence of genuine cognitive states.¹⁾

This author has chosen to focus the examination on the main ideas of one major critic in each of the three categories given by Casti, namely the arguments of John Searle, Hubert Dreyfus, and John Lucas. While this will have the unfortunate result of excluding many interesting and significant ideas of other AI researchers, this author hopes, thereby, to give the examination a degree of depth and clarity it would otherwise not have.

II. ORIGIN AND DEFINITION OF THE TERMS "ARTIFICIAL INTELLIGENCE"

Before proceeding, it would be good to look closely at the terms 'artificial intelligence' (AI), to better understand what computer scientists and researchers mean when they use these words. The words "artificial intelligence" were first used by Dartmouth mathematician John McCarthy at a summer study group held at Dartmouth in 1956. Sponsored by the Rockefeller Foundation, the alleged purpose of this study was to investigate "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it".²⁾

As can be seen here, the possibility of machine learning or machine thinking was seen, at that time, to be a "conjecture", i.e., a "guess" that it *might* be possible, based on little or no previously substantiated evidence. However, among the participants at this conference were a number of people who, optimistically believing in the possibility of AI, would go on to make the first serious attempts to make the conjecture a reality. Among them were, along with McCarthy, Marvin Minsky, who founded the AI Laboratory at M.I.T., Nobel Prize winner Herbert Simon, Claude Shannon, the "father" of information theory, Arthur Samuel, an early developer of game-playing programs, and others.

Of further note in the above statement are the notions of "precise description" and of "simulation". The first terms point to one of the fundamental assumptions held by some of the pioneers of AI that, be-

cause "intelligence" in humans was, in essence, the ability to represent reality in symbols and to manipulate these symbols according to the rules of logic, the means to accomplish intelligence artificially was to write sufficiently precise programs for computers to carry out. This idea was given formal expression by Newell and Simon as the "Physical Symbol System Hypothesis", as follows:

The digital computer has sufficient means for intelligent action, to wit: representing real-world objects, actions, and relationships internally as interconnected structures of symbols, and applying symbol manipulation to those structures.³⁾

Lenat and Feigenbaum refer to this hypothesis as "the founding principle of the AI research paradigm".⁴⁾

The term "simulation" indicates that there was some question as to the ontological status of such "artificial intelligence", should it actually be accomplished. Would such "intelligence" be identical with "natural intelligence" as manifested in and by human beings, in other words, a "duplication"; or would it somehow be "not quite the real thing", however similar, merely a "simulation". According to Lenat and Feigenbaum, the "physical symbol system hypotheses" as originally formulated by Newell and Simon was "an article of faith".⁵⁾ The ramifications of this distinction between "simulation" and "duplication" persists, according to Casti, even today, in the ways people think about the possibility of AI. These are examined in the next section.

III. FIVE DEGREES OF BELIEF

According to Jerome Shaffer: "Despite phenomenal progress in recent years, no computer yet devised even approximates in its capacity the multiplicitous powers of the human mind."⁶⁾ Thus, at this early stage in the development of AI, it is natural that people hold varying beliefs as to the possibility and type of AI which can be achieved. Casti gives a listing of five degrees of "strength", the origin of which he attributes to the philosopher Keith Gunderson:

--Strong AI, human: Whatever kinds of cognitive states machines might have, those states are functionally (although, of course, not physically) identical to those found in the human brain.

--Strong AI, nonhuman: The kinds of cognitive states found in a machine are not functionally identical to those in the brain and hence cannot be used to model human thought processes.

--Weak AI, sim-human: A computer can simulate human cognitive processes, but there is no particular correlation between the computer states and the cognitive states of the brain.

--Weak AI, sim-nonhuman: A computer can simulate the cognitive processes in a nonhuman mind (eg., a frog, a dog, an ant), but the states of the machine may or may not be re-

lated to those in the nonhuman brain.

--Weak AI, task, nonsim: The computer can perform tasks that previously required intelligence, but there is no intelligence required of the machine, whose states have nothing whatsoever to do with cognition, human or otherwise.⁷⁾

Casti goes on to state that,

As far as genuine machine thinking goes, the only category that counts is the first: strong AI, human; everything else, while undoubtedly technically challenging and economically rewarding, is pretty much devoid of any real intellectual or philosophical appeal, at least as far as the thinking-machine question goes.⁸⁾

The implication here is that those people who believe in the possibility of "strong AI, human" are asserting that they believe in the possibility that computers are capable of possessing cognitive states *functionally equivalent* and, hence, for all practical purposes, *identical* to those of human beings. At the very least, the advocates of "strong AI, human" would grant computers the ability to reason. The more daring would grant them the potential to manifest affective and volitional capacities (such as emotions and intentions), in addition to the more narrowly defined "cognitive" capacities, such as those involved in computation or logical reasoning. And there are some who would equate the possession of these mental capacities with "consciousness" or "awareness". Finally, at the furthest extreme in this category, are

those, such as Geoff Simons, author of a book entitled *Are Computers Alive ?*, and others, who plead the case for regarding computers as a newly emerging and evolving life form.

At first glance, the view(s) of the "strong AI, human" advocates strike one as contrary to the conventional views generally held by ordinary people about machines and other "inanimate" objects. Today's sophisticated machines, however, appear to possess "intentions", i.e., their "actions" and "behaviors" can be understood as being directed towards the accomplishment of set goals and purposes. But, as Shaffer writes:

The thesis that intentional phenomena are the essence of the mental...seems problematic. Its suggestion that the jet-searching missile has a mind or partakes of the mental...would, to many scholars, appear to be quite implausible. Nor does it seem that the trouble lies in the limited number of intentional phenomena found in the missile. Even the lunar-exploration machines, with all of their flexibility and multiplicity of functions, would not be said by most analysts to have minds.⁹⁾

Digital computers, however, seem to occupy a special place in the category of "machines", based on their unique ability to "process information". They are machines designed to deal flexibly with "intangibles", rather than to directly accomplish work in the physical world. In these respects, they seem to possess "mental powers" similar (i.e., functionally equivalent) to those of people, and naturally to suggest

that, with greater complexity and processing power designed into them, they would be capable of "intelligent behavior". Let us now look at some of the arguments which deny this possibility.

IV. "COMPUTERS CANNOT THINK": The Arguments against "Strong AI, Human"

At present, serious, heavily-funded research projects are taking place in many places around the world in the attempt to endow computers and robotic systems with "intelligence". Hence, it appears that the researchers, administrators, and others involved in these projects have accepted, a priori, the possibility of "strong AI, human". And if one listens to the media, current linguistic usage of the terms AI suggests that AI has already been largely achieved. But this is just the point at issue in the debate. The critics would argue just the opposite, or further, that AI is an impossibility using digital computers. Their arguments in opposition to the "strong AI, human" hypothesis derive from different sources, but each poses a thoughtful challenge to the hypothesis.

1. John Searle: "Syntax vs. Semantics"

John Searle is a philosopher in the Department of Philosophy at the University of California at Berkeley, whose published views in opposition to the "strong AI, human" hypothesis are very well-known to those familiar with the AI debate. According to Casti's classification above, Searle's arguments would fall into the last category: "anti-behavioristic arguments". In particular, his thought-experiment known as "the Chinese room" caused a sensational wave of reaction when it

was first published in 1981, and continues to hold, for this author, a degree of cogency derived from its appeal to logic and common sense. Although it has been widely referred to in the literature, I will once again repeat it here, in Searle's own words:

Imagine that a bunch of computer programmers have written a program that will enable a computer to simulate the understanding of Chinese. So, for example, if the computer is given a question in Chinese, it will match the question against its memory, or data base, and produce the appropriate answers to the questions in Chinese. Suppose for the sake of argument that the computer's answers are as good as those of a native Chinese speaker. Now then, does the computer, on the basis of this, understand Chinese, does it literally understand Chinese, in the way that Chinese speakers understand Chinese ?¹⁰⁾

Again, although common sense would suggest that the answer to the question posed by Searle should be "No", the advocates of the "strong AI, human" hypothesis would answer "Yes, the computer understands Chinese just as a native speaker does." Searle continues his parable in order to refute this conclusion, and in the process, the "strong AI, human" hypothesis as well:

Well, imagine that you are locked in a room, and in this room are several baskets full of Chinese symbols. Imagine that you (like me) do not understand a word of Chinese, but that you are given a rule book in English for manipulating

these Chinese symbols. The rules specify the manipulations of the symbols purely formally, in terms of their syntax, not their semantics.... Now suppose that some other Chinese symbols are passed into the room, and that you are given further rules for passing back Chinese symbols out of the room. Suppose that unknown to you the symbols passed into the room are called "questions" by the people outside the room, and the symbols you pass back out of the room are called "answers to the questions". Suppose, furthermore, that the programmers are so good at designing the programs and that you are so good at manipulating the symbols, that very soon your answers are indistinguishable from those of a native Chinese speaker.... On the basis of the situation as I have described it, there is no way you could learn any Chinese simply by manipulating these formal symbols.¹¹⁾

Here Searle is making an analogy to the way in which a digital computer operates by manipulating *formal symbols*, i.e., the carrying out of the internally stored program instructions on bits of data, which have syntax, but are completely lacking in semantic content, i.e., in meaning. Yet, according to the "physical symbol system hypothesis", the manipulation of "formal symbols" is "sufficient means for intelligent action". Searle would strongly disagree:

Now the point of the story is simply this: by virtue of implementing a formal computer program from the point of view of an outside observer, you behave exactly as if you understood Chinese, but all the same you don't understand a

word of Chinese. But if going through the appropriate computer program for understanding Chinese is not enough to give you an understanding of Chinese, then it is not enough to give any other digital computer an understanding of Chinese. And again, the reason for this can be stated quite simply. If you don't understand Chinese, then no other computer could understand Chinese, because no digital computer, just by virtue of running a program, has anything that you don't have. All that the computer has, as you have, is a formal program for manipulating Chinese symbols. *To repeat, a computer has a syntax, but no semantics.* The whole point of the parable of the Chinese room is to remind us of a fact that we knew all along. Understanding a language, or indeed, having mental states at all, involves more than just having a bunch of formal symbols. It involves having an interpretation, or a meaning attached to those symbols. And a digital computer, as defined, cannot have more than just formal symbols because the operation of the computer.... is defined in terms of its ability to implement programs. And these programs are purely formally specifiable—that is, they have no semantic content.¹²⁾ (*Italics mine*).

Searle's parable of the "Chinese room", by refuting the validity of the "physical symbol system hypothesis", strikes a blow at "the founding principle of the AI research paradigm". Searle summarizes his views opposing "strong AI, human" in a series of axioms and conclusion, as follows:

Axiom 1: Computer programs are formal (syntactic).

Axiom 2: Human minds have mental contents (semantics).

Axiom 3: Syntax by itself is neither constitutive of nor sufficient for semantics.

Conclusion 1: Programs are neither constitutive of nor sufficient for minds.¹³⁾

2. Hubert Dreyfus: "Nonprogrammable human capacities"

Hubert Dreyfus is currently a professor of Philosophy, at the University of California at Berkeley, where Searle also teaches. He first became interested in AI, however, when he was at M.I.T., because among his students were those who were concurrently enrolled in classes having to do with AI, and according to these students, solutions to many of the traditional philosophical issues concerning human mentality were being resolved by the achievements in computer science. Dreyfus was highly skeptical, and so began a long-lasting effort to refute the claims and question the assumptions of the AI researchers. Among his published works containing his arguments against the possibility of achieving AI using digital computers are: *What Computers Can't Do* (1972), revised as *What Computers Still Can't Do* in 1979 and 1992, and *Mind over Machine* (1989), written in collaboration with his brother, Stuart, also a professor at Berkeley in the Department of Industrial Engineering and Operations Research.

Specifically, Dreyfus focuses his lengthy and difficult arguments on two fundamental questions: "(1) Does a human being in "processing information" actually follow formal rules like a digital computer?", and (2) Can human behavior, no matter how generated, be described in a formalism which can be manipulated by a digital machine?"¹⁴⁾

Dreyfus bases his thoughtful skepticism in a philosophical perspective in which the ideas of the philosophers known as "phenomenologists" (Wittgenstein, Heidegger and Husserl, among others) play a prominent role:

In discussing each of these questions we found, first, that the descriptive or phenomenological evidence, *considered apart from traditional philosophical prejudices*, suggests that *nonprogrammable human capacities* are involved in all forms of intelligent behavior. Moreover, we say that no contrary empirical evidence stands up to methodological scrutiny. Thus, insofar as the question whether artificial intelligence is possible is an empirical question, the answer seems to be that further significant progress in Cognitive Simulation or in Artificial Intelligence is extremely unlikely.¹⁵⁾ (Italics mine).

Thus, in order to understand the main thrust of Dreyfus' position, it is necessary to first understand something about "phenomenology". The words I have italicized in the above passage are useful as a starting point, viz., that: 1) phenomenology presents a view of things different from what Dreyfus refers to as "traditional philosophical prejudices"; and 2) phenomenology argues against the validity of the "physical symbol system hypothesis". In 1), above, Dreyfus is here referring to the ideas of the Western philosophical tradition going back to Plato:

Since the Greeks invented logic and geometry, the idea that all reasoning might be reduced to some kind of calculation—so

that all arguments could be settled once and for all—have fascinated most of the Western tradition's rigorous thinkers. Socrates was the first to give voice to this vision. The story of artificial intelligence might well begin around 450 B.C. when (according to Plato) Socrates demands of Euthyphro, a fellow Athenian who, in the name of piety, is about to turn in his own father for murder: "I want to know what is characteristic of piety which makes all actions pious, that I may have it to turn to, and to use as a standard whereby to judge your actions and those of other men." Socrates is asking Euthyphro for what modern computer theorists would call an "effective procedure," "a set of rules which tells us, from moment to moment, precisely how to behave."¹⁶⁾

Dreyfus continues:

For the Platonic project to reach fulfillment one breakthrough is required: *all appeal to intuition and judgement must be eliminated*. As Galileo discovered that one could find a pure formalism for describing physical motion by ignoring secondary qualities and teleological considerations, so, one might suppose, a Galileo of human behavior might succeed in reducing all semantic considerations (appeal to meanings) to the techniques of syntactic (formal) manipulation. *The belief that such a total formalization of knowledge must be possible came to dominate Western thought*. It already expressed a basic moral and intellectual demand, and the success of physical science seemed to imply to sixteenth-century philo-

sophers, as it still seems to suggests to thinkers such as Minsky, that the demand could be satisfied.¹⁷⁾ (*Italics mine*).

As we saw above in Searle's critique, the "strong AI, human" hypothesis is founded on just such an idea, to which Newell and Simon gave the name the "physical symbol system hypothesis", and the implication of the above passage is that, in Dreyfus' view, such a belief derives from "traditional philosophical prejudices" that have allowed AI researchers to over-optimistically assume the possibility of achieving artificial intelligence through syntactic symbol manipulation by digital computers. As Searle concisely puts it in his "axiom 3": "Syntax by its itself is neither constitutive of nor sufficient for semantics."

But Dreyfus goes well beyond the behavioristic arguments of Searle in his detailed, phenomenological critique of what he sees as four specific assumptions underlying the general assumption on the part of AI and CS (Cognitive Simulation) workers that "man functions like a general-purpose symbol-manipulating device".¹⁸⁾ The four underlying assumptions are:

1. A biological assumption that on some level of operation—usually supposed to be that of the neurons—the brain processes information in discrete operations by way of some biological equivalent of on/off switches.
2. A psychological assumption that the mind can be viewed as a device operating on bits of information according to for-

mal rules. Thus, in psychology, the computer serves as a model of the mind...[a] model of thinking as data processing—a third-person process in which the involvements of the “processor” plays no essential role.

3. An epistemological assumption that all knowledge can be formalized, that is, that whatever can be understood can be expressed in terms of logical relations, more exactly in terms of Boolean functions, the logical calculus which governs the way the bits are related according to rules.

4. Finally since all information fed into digital computers must be in bits, the computer model of the mind presupposes that all relevant information about the world, everything essential to the production of intelligent behavior, must in principle be analyzable as a set of situation-free determinate elements. This is the ontological assumption that what there is, is a set of facts each logically independent of all the others.¹⁹⁾

With reference to these four assumptions, Dreyfus states that each

...assumption is taken by workers in CS (Cognitive Simulation) or AI as an axiom, guaranteeing results, whereas it is, in fact, only one possible hypothesis among others, to be tested by the success of such work. Furthermore, none of the four assumptions is justified on the basis of the empirical and *a priori* arguments brought forward in its favor. Finally,

the last three assumptions, which are philosophical rather than empirical, can be criticized on philosophical grounds. They each lead to conceptual difficulties when followed through consistently as an account of intelligent behavior.²⁰⁾

In refutation of the biological assumption, Dreyfus argues that the idea that "the brain processes information in discrete operations by way of some biological equivalent of on/off switches" is an outmoded one based on the early discoveries by neurophysiologists of the apparent "all-or-none firing behavior" of neurons. The analogy to the on/off operations of digital switches is a natural one to make, but the leap to the conclusion that the brain, as a whole, operates like a digital processor has been called into question by later empirical discoveries that suggest that the brain operates more like an analogue computer, with massively parallel processing of decentralized, highly distributed "data", and with individual neurons responding to an algebraic summation of incoming impulses. He concludes his argument:

Thus the view that the brain as a general-purpose symbol manipulating device operates like a digital computer is an empirical hypothesis which has had its day. No arguments as to the possibility of artificial intelligence can be drawn from the current empirical evidence concerning the brain. In fact, the difference between the "strongly interactive" nature of brain organization and the noninteractive character of machine organization suggest that insofar as arguments from biology are relevant, the evidence is against the possibility of using digital computers to produce intelligence.²¹⁾

The psychological assumption underlies a theory of how *the mind* functions, as opposed to how *the brain* functions, i.e., that, at a certain level of operation "above" the biological level, the mind *processes information*, and does so in ways similar to how a digital computer operates by "comparing, classifying, searching lists, and so forth, to produce intelligent behavior".²²⁾

Dreyfus argues:

...that the assumption of an information-processing level is by no means so self-evident as the cognitive simulators seem to think; that there are good reasons to doubt that there is any information processing going on, and therefore reason to doubt the validity of the claim that the mind functions like a digital computer.²³⁾

Among the reasons he gives is to point out confusion in the meaning and usage of the term "information processing":

"Information processing" is ambiguous. If this term simply means that the mind takes account of meaningful data and transforms them into other meaningful data, this is certainly incontrovertible. But the cybernetic theory of information, introduced in 1948 by Claude Shannon, has nothing to do with meaning in this ordinary sense. It is a nonsemantic, mathematical theory of the capacity of communication channels to transmit data. ...

When illegitimately transformed into a theory of mean-

ing, in spite of Shannon's warning, information theory and its vocabulary have already built in the computer-influenced assumption that experience can be analyzed into isolable, atomic, alternative choices. As a theory of meaning this assumption is by no means obvious. ... Much of the literature of Cognitive Simulation gains its plausibility by shifting between the ordinary use of the term "information" and the technical sense the term has recently acquired... we must be careful to speak and think of "information processing" in quotation marks when referring to human beings.²⁴⁾

As further reason for urging this caution, Dreyfus takes a phenomenological viewpoint:

Gestalt psychologists... claim that thinking and perception involve global processes which cannot be understood in terms of a sequence or even a parallel set of discrete operations. Just as the brain seems to be, at least in part, an analogue computer, so the mind may well arrive at its thoughts and perceptions by responding to "fields," "forces," "configurations," and so on, as, in fact, we seem to do insofar as our thinking is open to phenomenological description.²⁵⁾

He argues further that:

Moreover, even if the mind did process information in Shannon's sense of the term, and thus function like a digital computer, there is no reason to suppose that it need to so

according to a program. If the brain were a network of randomly connected neurons, there might be no flow chart, no series of rule-governed steps on the information-processing level, which would describe its activity.

Both these confusions--the step from ordinary meaning to the technical sense of information and from computer to heuristically programmed digital computer--are involved in the fallacy of moving from the fact that the brain in some sense transforms its inputs to the conclusion that the brain or mind performs some sequence of discrete operations.²⁶⁾

Dreyfus claims that both CS and AI workers (Newell, Neisser, Miller, and Fodor) accept the psychological assumption *a priori*, i.e., based on a theory (of how the mind operates), rather than on empirical evidence. His arguments designed to prove that there is no valid justification for making the psychological assumption are, unfortunately, too technical and detailed for the scope of this paper, so I refer the reader to Dreyfus' book, *What Computers Still Can't Do*, pp. 174-187. In his refutation regarding the validity of the psychological assumption, Dreyfus concludes:

The answer to the question whether man can make [an intelligent] machine must rest on the evidence of work being done. And on the basis of actual achievements and current stagnation, the most plausible answer seems to be, No...although man is surely a physical object processing physical inputs according to the laws of physics and chemistry, man's behavior may not be explainable in terms of an information-

processing mechanism processing inputs which represent features of the world. Nothing from physics or experience suggests that man's actions can be so explained, since on the physical level we are confronted with continuously changing patterns of energy, and on the phenomenological level with objects in an already organized field of experience.²⁷⁾

The epistemological assumption is that "although human performance might not be *explainable* by supposing that people are actually following heuristic rules in a sequence of unconscious operations, intelligent behavior may still be *formalizable* in terms of such rules and thus reproduced by machine."²⁸⁾ Dreyfus refers to the success of formalization in the physical sciences (Galileo, et al.) and linguistics (Chomsky) as the basis for the optimism of AI workers, such as Minsky.

Regarding the former, however, Dreyfus claims that using the success of formalizations in describing natural phenomenon, eg., the motions of the planets, as the basis for claiming that "all nonarbitrary behavior can be formalized", including that of human behavior, is an "unjustified generalization".²⁹⁾

In the case of the latter, while Dreyfus grants that Chomsky and his followers have discovered the basis for a theory (i.e., a formalization) of linguistic *competence* in humans (i.e. their ability to "recognize grammatically well-formed sentences and to reject ill-formed ones").³⁰⁾ such theory is insufficient to explain actual linguistic *performance*, as such performance does not appear to depend upon the obeying of

rules by human speakers. As evidence of this, Dreyfus points out that humans demonstrate linguistic comprehension of “odd” or “contrary-to-rule” linguistic expressions, based on context:

“The idea is in the pen” is clear in a situation in which we are discussing promising authors; but a machine at this point, with rules for what size physical objects can be in pig pens, playpens, and fountain pens, would not be able to go on. Since an idea is not a physical object, the machine could only deny that it could be in the pen or at best make an arbitrary stab at interpretation. The listener’s understanding, on the other hand, is far from arbitrary. Knowing what he does about the shadow which often falls between human projects and their execution, as well as what one uses to write books, he gets the point, and the speaker will often agree on the basis of the listener’s response that the listener has understood. Does it follow, then, that in understanding or using the odd utterance, the human speakers were acting according to a rule—in this case a rule for how to modify the meaning of “in”? It certainly does not seem so to the speakers who have just recognized the utterance as “odd.”³¹⁾

Dreyfus, following Wittengenstein, claims that if it were necessary for humans to obey rules in order to use language (performance), there would have to be “meta-rules” in order to guide speakers and listeners in applying these rules and in interpreting and understanding the exceptions to the rules, and “meta-meta-rules” for applying the “meta-rules”, and so on, in an infinite regress:

To have a complete theory of what speakers are able to do, one must not only have grammatical and semantic rules but further rules which would enable a person or a machine to recognize the context in which the rules must be applied. Thus there must be rules for recognizing the situation, the intentions of the speakers, and so forth. But if the theory then require further rules in order to explain how these rules are applied...we are in an infinite regress. Since we do manage to use language, this regress cannot be a problem for human beings. If AI is to be possible, it must also not be a problem for machines.³²⁾

Human beings escape from this infinite regress because "[A]t some level...the interpretation of the rule is simply evident and the regress stops."³³⁾ Likewise, for machines,

but interpretation *has nothing to do with the demands of the situation. It cannot, for the computer is not in a situation.* It generates no local content. The computer theorist's solution is to build the machine to respond to *ultimate bits of context-free, completely determinate data* which require no further interpretation in order to be understood...as, for example, holes in cards or the mosaic of a TV camera, so on this ultimate level the machine does not need rules for applying its rules.³⁴⁾ (Italics mine).

In the phenomenologists' view, the behavior of human beings (in

contrast to that of computers) is always "rooted in a situation" in which there are no "ultimate bits of context-free, completely determinate data" on which to base an interpretation that guides the application of rules and meta-rules. Hence, Dreyfus would argue, the epistemological assumption does not provide a sound basis for a theory of human psychology which would equate human and machine performance. This objection, however, being "based on appearances", does not suffice, Dreyfus feels, to convince "those committed to the epistemological assumption":

A full refutation of the epistemological assumption would require an argument that the world *cannot* be analyzed in terms of context-free data. Then, since the assumption that there are basic unambiguous elements is the only way to save the epistemological assumption from the regress of rules, the formalist, caught between the impossibility of always having rules for the application of rules and the impossibility of finding ultimate unambiguous data, would have to abandon the epistemological assumption altogether.³⁵⁾

The epistemological assumption is thus seen to rest upon another, which Dreyfus calls "the ontological assumption" regarding the ultimate nature of the world as analyzable into context-free data, and which Dreyfus says is "the deepest assumption underlying work in AI and the whole philosophical tradition"³⁶⁾ Dreyfus writes:

...we have seen that the biological, psychological, and epistemological assumptions...are totally unjustified and may well

be untenable. Now we turn to an even more fundamental difficulty facing those who hope to use digital computers to produce artificial intelligence: the data with which the computer must operate if it is to perceive, speak, and in general behave intelligently, must be *discrete, explicit, and determinate*; otherwise, it will not be the sort of information which can be given to the computer so as to be processed by rule. Yet there is no reason to suppose that such data about the human world are available to the computer and several reasons to suggest that no such data exist. (Italics mine).

The ontological assumption that everything essential to intelligent behavior must in principle be understandable in terms of a set of determinate independent elements...lies at the basis of all thinking in AI, and...it can seem so self-evident that it is never made explicit or questioned...Once this hypothesis is made explicit and called into question, it turns out that no arguments have been brought forward in its defense and that, when used as the basis for a theory of practice such as AI, the ontological assumptions leads to profound conceptual difficulties.³⁷⁾

AI pioneer Marvin Minsky once estimated that "a machine will quite critically need to acquire the order of a hundred thousand elements of knowledge in order to behave with reasonable sensibility in ordinary situations. A million, if properly organized, should be enough for a very great intelligence".³⁸⁾ Whatever the true number of "elements of knowledge" required for intelligence, and ignoring the practical problems of designing a computer system capable of acces-

sing such knowledge in a reasonable amount of time (the "large data base problem"), Dreyfus implies that even in simply making such an estimate, Minsky makes the ontological assumption, which Dreyfus feels lacks justification:

It is by no means obvious that in order to be intelligent, human beings have somehow solved or needed to solve the large data base problem. The problem may itself be an artifact created by the fact that AI workers must operate with discrete elements. Human knowledge does not seem to be analyzable as an explicit description as Minsky would like to believe. A mistake, a collision, an embarrassing situation, etc. do not seem on the face of it to be objects or facts about objects. Even a chair is not understandable in terms of any set of facts or "elements of knowledge". To recognize an object as a chair, for example, means to understand its relations to other objects and to human beings. This involves *a whole context of human activity* of which the shape of our body, the institution of furniture, the inevitability of fatigue, constitute only a small part. And *these factors in turn are no more isolable than is the chair. They all may get their meaning in the context of human activity of which they form a part.*³⁹⁾ (Italics mine).

The basic problem, according to Dreyfus, is that "in order to understand an utterance, structure a problem, or recognize a pattern, a computer must select and interpret its data in terms of a context. But how are we to impart this context itself to the computer?"⁴⁰⁾ If

the ontological assumption is unjustifiable, i.e., if the human world of phenomena is ultimately not analyzable into "context-free-data", or "elements of knowledge", then it is not possible to provide the computer with the required, relevant context.

Dreyfus again takes the problem of the ambiguity of natural language to illustrate his argument, and refers to Joseph Weizenbaum's proposal "to program a nest of contexts in terms of a "contextual tree":⁴¹⁾ "beginning with the topmost or initial node, a new node representing a subcontext is generated, and from this one a new node still, and so on to many levels"⁴²⁾ as a way to provide the computer with the necessary context for "disambiguating" natural language:

To understand why Weizenbaum finds it necessary to use a hierarchy of contexts and work down from the top node, we must return to the general problem of situation recognition. If computers must utilize the situation or context in order to disambiguate, and in general to understand utterances in a natural language, the programmer must be able to program into the machine, which is not involved in a situation, a way of recognizing a context and using it. But the same two problems which arose in disambiguation and necessitated appeal to the situation in the first place arise again on the level of context recognition and force us *to envisage working down from the broadest context*: (1) If in disambiguation the number of possibly relevant facts is in some sense infinite so that selection criteria must be applied before interpretation can begin, the number of facts that might

be relevant to recognizing a context is infinite too. How is the computer to consider all the features such as how many people are present, the temperature, the pressure, the day of the week, and so forth, any one of which might be a defining feature of some context? (2) Even if the program provides rules for determining relevant facts, the facts would be ambiguous, that is, capable of defining several different contexts, until they were interpreted.

Evidently a broader context will have to be used to determine which of the infinity of features is relevant, and how each is to be understood. But if, in turn, the program must enable the machine to identify the broader context in terms of its relevant features...the programmer must either claim that some features are intrinsically relevant and have a fixed meaning regardless of context—a possibility already excluded in the original appeal to context—or the programmer will be faced with an infinite regress of contexts. There seems to be only one way out: rather than work up the tree to ever broader contexts, the computer must work down from an ultimate context—what Weizenbaum calls our shared culture.⁴³⁾

What Weizenbaum calls “our shared culture”, Dreyfus calls the “context of social intercourse”; yet even this broad context can be viewed as a subcontext of “human activity”, which is, in turn, a subcontext of what Dreyfus calls the “human life-world” (what Wittgenstein, with reference to the ultimate situation of people referred to as “forms of life”). Thus, the computer programmer is faced with an infi-

nite regress of broader and broader contexts necessary for providing a basis for determining the relevance, hence the meaning, of facts, utterances, and words. Dreyfus ponders: "Well then, why not make explicit the significant features of the human form of life from within it?"⁴⁴⁾ But he immediately replies:

Indeed, this *deus ex machina* solution has been the implicit goal of philosophers for two thousand years, and it should be no surprise that nothing short of a formalization of the form of life could give us artificial intelligence (which is not to say that this is what gives us normal intelligence). But how are we to proceed? Everything we experience in some way, immediate or remote, reflects our human concerns. Without some *particular* interest, without some *particular* inquiry to help us select and interpret, we are back confronting the infinity of meaningless facts we were trying to avoid.

It seems that given the artificial intelligence worker's conception of reason as calculation on facts, and his admission that which facts are relevant and significant is not just given but is context determined, his attempt to produce intelligent behavior leads to an antinomy. On the one hand, we have the thesis: there must always be a broader context; otherwise, we have no way to distinguish relevant from irrelevant facts. On the other hand, we have the antithesis: there must be an ultimate context, which requires no interpretation; otherwise, there will be an infinite regress of contexts and we can never begin our formalization.⁴⁵⁾

Fortunately, in contrast to computers, we human beings seem to “embody a third possibility which would offer a way out of this dilemma”, namely the ability to recognize the present situation “as a continuation or modification of the previous one”, by carrying over “from the immediate past a set of anticipations based on what was relevant and important a moment ago”.⁴⁶⁾ But how humans accomplish this task is still not well-understood.

In the final part of his analysis of the four assumptions, Dreyfus concludes:

In surveying the four assumptions underlying the optimistic interpretation of results in AI, we have observed a recurrent pattern: In each case the assumption was taken to be self-evident—an axiom seldom articulated and never called into question. In fact, the assumption turned out to be only one alternative hypothesis, and a questionable one at that. The biological assumption that the brain must function like digital computer no longer fits the evidence. The others lead to conceptual difficulties.

The psychological assumption that the mind must obey a heuristic program cannot be defended on empirical grounds, and a priori arguments in its defense fail to introduce a coherent level of discourse between the physical and the phenomenological. ...

[The] fundamental difficulty [of how to formalize the totality of human knowledge presupposed in intelligent behavior] is hidden by the epistemological and ontological

assumptions that all human behavior must be analyzable in terms of rules relating atomic facts.

But the conceptual difficulties introduced by these assumptions are even more serious than those introduced by the psychological one. The inevitable appeal to these assumptions as a final basis for a theory of practice leads to a regress of more and more specific rules for applying rules or of more and more general contexts for recognizing contexts. In the face of these contradictions, it seems reasonable to claim that, *on the information processing level, as opposed to the level of the laws of physics, we cannot analyze human behavior in terms of rule-governed manipulation of a set of elements*. And since we have seen no argument brought forward by the AI theorists for the assumption that human behavior must be reproducible by a digital computer operating with strict rules on determinate bits, we would seem to have good philosophical grounds for rejecting this assumption.⁴⁷⁾ (Italics mine).

3. John R. Lucas: The mind always has "the last word"

In his now-famous 1950 paper entitled "Computing Machinery and Intelligence", Alan Turing anticipated arguments opposing the idea of artificial intelligence based on mathematics, in particular those which looked to the implications of Gödel's proof of the inherent limitations of formal systems as a basis for hailing the superiority of humans over machines. Turing wrote:

There are a number of results of mathematical logic which

can be used to show that there are limitations to the powers of discrete state machines. The best known of these results is known as Gödel's theorem, and shows that in any sufficiently powerful logical system statements can be formulated which can neither be proved nor disproved within the system, unless possibly the system itself is inconsistent.⁴⁸⁾

To see how it is possible to make such an assertion, it is first necessary to understand (at least) the "results" (if not the detailed method) of Gödel's proof,⁴⁹⁾ which pertain to the fundamental impossibility of any formal system to be *simultaneously* both *consistent* and *complete* (i.e., "consistent" in the sense of not being able to produce logically antithetical statements such as "A" and "not-A", and "complete" in the sense of being able to provide proofs of all possible legitimate formulae derivable within the system), and the fundamental impossibility of a formal system to *supply proof of its own consistency*, and to realize, at the same time, that "discrete state machines" (Turing's name for "digital computers", which, at the time he wrote his paper, were just in the process of becoming a reality) are *equivalent to formal systems*. Hence, the limitations inherent in such systems are likewise inherent in digital computers. Therefore, if one could rightly claim that human beings do not suffer from such limitations, one could argue that digital computers are not capable of achieving true, human-being-like intelligence. Turing replies:

The short answer to this argument is that although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort

of proof, that no such limitations apply to the human intellect....We too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines. Further, our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph. There would be no question of triumphing simultaneously over all machines. In short, then, there might be men cleverer than any given machine, but then again there might be other machines cleverer again, and so on.⁵⁰⁾

British philosopher John R. Lucas attempted to refute Turing's view in his 1961 paper entitled "Minds, Machines, and Gödel":

However complicated a machine we construct, it will, if it is a machine, correspond to a formal system, which in turn will be liable to the Gödel procedure for finding a formula unprovable-in-that-system. This formula the machine will be unable to produce as being true, although a mind can see it is true. And so the machine will still not be an adequate model of the mind. We are trying to produce a model of the mind which is mechanical—which is essentially "dead"—but the mind, being in fact "alive," can always go one better than any formal, ossified, dead system can. Thanks to Gödel's theorem, the mind always has the last word.⁵¹⁾

In reply to Lucas argument, Douglas R. Hofstadter has written:

On first sight, and perhaps even on careful analysis, Lucas' argument appears compelling. It usually evokes rather polarized reactions. Some seize onto it as a nearly religious proof of the existence of souls, while others laugh it off as being unworthy of comment. I feel it is wrong, but fascinatingly so—and therefore quite worthwhile taking the time to rebut....

We must try to understand more deeply why Lucas says the computer cannot be programmed to “know” as much as we do. Basically the idea is that we are always *outside* the system, and from out there we can always perform the “Gödelizing” operation, which yields something which the program, from within, can't see is true.⁵²⁾

The obscure terms “Gödelizing operation”, which Hofstadter borrows from Lucas, refers to derivation of a particular theorem (which Lucas calls “the Gödelian formula”) for any given formal system and then adding this theorem to the axiom set of the system, and doing so again and again, in an infinite sequence, in a futile attempt to achieve “completeness” within the formal system. The “Gödelian formula”, derived by applying the logical techniques developed by Gödel, has the special property of “unprovability” within the formal system to which it refers. We human beings, being outside the system, can see that the Gödelian formula is a “true formula”, but the formal system embodied in a computer programmed to perform the “Gödelian operation” cannot do so. Thus, as Lucas puts it:

...We might expect a mind, faced with a machine that possessed

a Gödelizing operator, to take this into account, and out-Gödel the new machine, Gödelizing operator and all. This has, in fact proved to be the case. Even if we adjoin to a formal system the infinite set of axioms consisting of the successive Gödelian formulae, the resulting system is still incomplete, and contains a formula which cannot be proved-in-the-system, although a rational being can, standing outside the system, see that it is true. We had expected this, for even if an infinite set of axioms were added, they would have to be specified by some finite rule or specification, and this further rule or specification could then be taken into account by a mind considering the enlarged formal system. In a sense, just because the mind has the last word, it can always pick a hole in any formal system presented to it as a model of its own workings. The mechanical model must be, in some sense, finite and definite: and then the mind can always go one better.⁵³⁾

Hofstadter takes issue with Lucas on the point that humans can do something which computers cannot:

...the very fact that we cannot write a program to do "Gödelizing" must make us somewhat suspicious that we ourselves could do it in every case. It is one thing to make the argument in the abstract that Gödelizing "can be done"; it is another thing to know how to do it in every particular case. In fact, as the formal systems (or programs) escalate in complexity, our own ability to "Gödelize" will eventually begin

to waver. It must, since...we do *not* have any algorithmic way of describing how to perform it. If we can't tell *explicitly* what is involved in applying the Gödel method in all cases, then for each of us there will eventually come some case so complicated that we simply can't figure out how to apply it.⁵⁴⁾

V. Rebutting the Critics: Casti

In any good debate, there will be a lively, if not outright hostile, exchange of differing views, with claims and counter-claims, assertions and rebuttals. Such is also the case with the "AI Debate". In his chapter entitled "The Cognitive Engine", Casti, after presenting his own versions of the arguments of the three critics discussed above, concludes with his own evaluation of their merits and demerits, and with a statement of his personal belief in the possibility of "strong AI, human".

In order to fully understand Casti's refutations, however, it is necessary to introduce a further distinction regarding the basis for human intelligence, viz., the distinction between "top-down" and "bottom-up" views of the nature of intelligence. Briefly, the "top-down" view is that intelligence consists of "higher order" *cognitive functions*, such as logical reasoning, sorting, listing, use of language, symbolic representations of reality manipulated by rules, etc. These are just the sorts of functions that can be carried out by suitably programmed digital computers. In the "top-down" view, intelligence is more a matter of "software", with the type of "hardware" being irrelevant. The "bottom-up" view takes an opposite approach, i.e., that intelligence is

an "*epiphenomenon*" that "*emerges*" from the unintelligent physical and chemical activities of the brain on the lowest level, and hence, hardware matters. An intelligent computer would need to bear a close similarity to *the physical structure of the brain* prior to its having the ability to demonstrate mental capacity.

In reverse order, let's look at Casti's replies to the critics. Regarding Lucas, Casti writes that his

appeal to Gödel has the surface ring of something you can really get your teeth into: tangible, to the point, and mathematically airtight. But Gödel's results, like all high-precision tools, apply to a very definite and severely restricted set of circumstances, and it seems to me that Lucas has stretched these conditions beyond the breaking point in his effort to invoke Gödel as an argument against thinking machines...Gödel himself didn't appear to see his work as any kind of obstacle to the existence of intelligent machines. And what's good enough for Gödel is certainly good enough for me! Thus does Lucas...fall by the wayside.⁵⁵⁾

Next, Casti feels that it is "easy to drop the Dreyfus argument from the list of contenders", because:

...The core of the Dreyfuses' claim is the phenomenological assertion that many crucial aspects of human thinking like judgement, understanding, and perception cannot be formalized...The whole edifice of the Dreyfus case rests on what amounts to the religious claim that Husserl, Heidegger,

& Co. are right. But to my eye, the Dreyfuses put forth anything but a knockdown argument supporting this crucial assumption. Furthermore, I think it's important to note that they are primarily arguing against the top-down AI programs of the Simon and Newell type. Thus, even if through some unforeseeable set of circumstances their phenomenological thesis could be proven correct, I fail to see how this fact would begin to touch upon the program of the bottom-up school. Putting these observations together, I think it's also safe to scratch the Dreyfuses from the race.⁵⁶⁾

Finally, Casti writes:

Oddly enough, I find Searle's arguments based on the first-person perspective of the Chinese Room to be the most compelling, and it's with some trepidation that I finally cast it aside along with the others. The two axioms underpinning Searle's claims are (1) brains cause mental states, and (2) no amount of syntax alone can ever generate semantics; i.e., no amount of form will ever produce content or meaning. Personally, I have reservations about the first point and completely disagree with the second.⁵⁷⁾

VI. Conclusion

And so it goes, argument and counter-argument, with neither side in the debate succeeding in convincing the opposing side of the rightness of its point of view. But, *the debate has merit in itself in its caus-*

ing us to take a closer look at the issues, and by bringing the various difficulties, philosophical and practical, into sharper focus.

In conclusion, this author is fully in agreement with the sentiments of Casti when he writes, in summary:

On balance, it seems to me that the thinking-machine debate is really a battle between philosophers, regardless of the fact that some of them may be masquerading as psychologists, computer scientists, mathematicians, or programmers. And, as it should be in all stories involving philosophers, the debate ends up in complete chaos. My gut feeling is that a genuine machine intelligence will be with us within the next decade or two, but I'll have to confess that the opinion is based as much upon wishing, hoping and wondering as upon hard facts and philosophical arguments.... However the matter of strong AI, human, is finally resolved, the outcome will radically change our view of ourselves and our perception of the place we occupy in the cosmic order of things.⁵⁸⁾

In this study this author has adopted Casti's scheme for categorizing the types of arguments put forth against the "strong AI, human" hypothesis, viz., *antibehavioristic* (Searle), *phenomenological* (Dreyfus), and *logical* (Lucas). Obviously, there are many other thinkers whose ideas regarding AI, pro and con, deserve serious consideration. In forthcoming papers, this author intends to introduce the claims and achievements of those who have all along believed in the possibility of AI, and to look at the directions in which the AI enterprise is headed.

VII. Endnotes

- 1) Casti, John L., *Paradigms Lost: Tackling the Unanswered Mysteries of Modern Science*, Avon Books, New York, 1989, p. 315.
- 2) as quoted in Casti, John L., *Paradigms Lost*, p. 285.
- 3) in Lenat, D.B. and E.A. Feigenbaum, "On the thresholds of knowledge", in David Kirsh, ed. *Foundations of Artificial Intelligence*, MIT/Elsevier, 1992, pp. 193-194.
- 4) Ibid, p. 193.
- 5) Ibid.
- 6) Shaffer, Jerome A., in "Mind, The Philosophy of", *Encyclopedia Britannica*, p. 125.
- 7) Casti, *Paradigms Lost*, p. 286.
- 8) Ibid., p. 287.
- 9) Shaffer, in "Mind, The Philosophy of", p. 119.
- 10) Searle, John, *Minds, Brains and Science*, Harvard University Press, 1984, p. 32.
- 11) Ibid., p. 32.
- 12) Ibid., p. 33.
- 13) Searle, John R., "Is the Brain's Mind a Computer Program ?", in *Scientific American*, January 1990, Volume 262, Number 1, p. 27.
- 14) Dreyfus, Hubert, *What Computers Still Can't Do: A Critique of Artificial Reason*, M.I.T. Press, 1992, p. 285.
- 15) Ibid., p. 285.
- 16) Ibid., p. 67.
- 17) Ibid., p. 69.
- 18) Ibid., p. 156.
- 19) Ibid.
- 20) Ibid., p. 157.
- 21) Ibid., p. 162.
- 22) Ibid., p. 163.
- 23) Ibid.
- 24) Ibid., pp. 165-166.
- 25) Ibid., p. 166.
- 26) Ibid.
- 27) Ibid., pp. 187-188.
- 28) Ibid., p. 189.
- 29) Ibid., p. 190.
- 30) Ibid., p. 198.
- 31) Ibid., p. 199.
- 32) Ibid., p. 203-204.
- 33) Ibid., p. 204.

- 34) Ibid.
- 35) Ibid., p. 205.
- 36) Ibid.
- 37) Ibid., p. 206-207.
- 38) Minsky, Marvin, ed., *Semantic Information Processing*, M.I.T. Press, 1969, p. 25, as quoted in Dreyfus, *What Computers Still Can't Do*, p. 209.
- 39) Dreyfus, *What Computers Still Can't Do*, p. 210.
- 40) Ibid., p. 208.
- 41) Ibid., p. 219.
- 42) Joseph Weizenbaum, "Contextual Understanding by Computers," in Kolers and Eden, eds., *Recognizing Patterns: Studies in Living and Automatic Systems*, M.I.T. Press, 1968, pp. 181-182, in Dreyfus, *What Computers Still Can't Do*, p. 219.
- 43) Dreyfus, *What Computers Still Can't Do*, pp. 220-221.
- 44) Ibid., p. 221.
- 45) Ibid., p. 222.
- 46) Ibid.
- 47) Ibid., p. 225-226.
- 48) Turing, Alan, "Computing Machinery and Intelligence", in Douglas R. Hofstadter and Daniel C. Dennett, eds., *The Mind's I*, Bantam Books, 1981, p. 58.
- 49) For readers interested in understanding the details of the proof, the following books are recommended: 1) *Gödel's Proof* by Ernest Nagel and James R. Newman (New York University Press, 1958) and *Gödel, Esher, Bach: An Eternal Golden Braid* by Douglas R. Hofstadter (Penguin Books, 1979).
- 50) Ibid., p. 59.
- 51) Lucas, John R., "Minds, Machines, and Gödel", in Alan Ross Anderson, ed., *Minds and Machines*, Prentice Hall, 1964, p. 43, in Douglas R. Hofstadter, *Gödel, Esher, Bach: An Eternal Golden Braid*, Penguin Books, 1980, p. 472.
- 52) Hofstadter, *Gödel, Esher, Bach*, p. 472.
- 53) Lucas, John R., "Minds, Machines, and Gödel", in Alan Ross Anderson, ed., *Minds and Machines*, Prentice Hall, 1964, p. 48-49, in Douglas R. Hofstadter, *Gödel, Esher, Bach: An Eternal Golden Braid*, Penguin Books, 1980, p. 473.
- 54) Hofstadter, *Gödel, Esher, Bach*, p. 475.
- 55) Casti, *Paradigms Lost*, p. 333.
- 56) Ibid.
- 57) Ibid., p. 334.
- 58) Ibid., p. 339.

AI 論争：疑問の声

この論文で、AI論争においていわゆる反対陣営の交わす反論の哲学的内容を調べてみた。AIの実現にあたって次々に出てくる難問に対する理解を深めるのにはこの反論は重要な役割を果たしている。John Casti は AI の概念に対する反論の哲学的内容を三つの題目にまとめている。即ち、反行動的反論、現象学的反論、そして論理的反論である。これは、人間の頭脳にみられる認識状態と同等な状態を備えた機械（計算器）が理論的に作れるとする高度な人間的機能を持つ AI 理論に対して批判的な立場にある哲学者の主張である。ここで、各々の反論派の代表的論者の意見を調べてみた。反行動的反論の場合はカリフォルニア大学、バークレー校の John Searle が提唱した有名な Chinese Room Thought Experiment、現象学的反論の場合は同じバークレー校の Hubert Dreyfus の主張、そして論理的反論の場合はよく知られている Gödel's Proof に基づくイギリスの John Lucas の主張である。

歴史的にみると、artificial intelligence という用語は Dartmouth 大学の数学者の John McCarthy が1956年、夏期研修グループで初めて使っている。Rockefeller Foundation 主催のこの研修会の主な課題は、「学問または知能のあらゆる様相は、原則として機械によってシミュレーションできるほど正確に記述することができる。」という仮定が立証できるかどうかということであった。

Casti は上記の三種の反論には納得していない。Casti は真の意味での知的な機械は二、三十年以内に開発できると確信している。著者も Casti の意見に同意している。AI を巡る論争はこれからも続くが、論争は問題をより深く追求する機会を与える一方 AI に関わる哲学的また実践的な諸課題をより明確にするであろう。