Hilary Putnam

Much Ado About Not Very Much

THE QUESTION I WANT TO CONTEMPLATE is this: Has artificial intelligence taught us anything of importance about the mind? I am inclined to think that the answer is no. I am also inclined to wonder, What is all the fuss about? Of course, AI may someday teach us something important about how we think, but why are we so exercised now? Perhaps it is this prospect that exercises us, but why do we think now is the time to decide what might in principle be possible? Or am I wrong: Is the "in principle" question really the important one to discuss now? And if it is, have the defenders of AI had anything important to tell us about it?

The computer model of the mind is now associated with AI, but it is not unique to AI (Noam Chomsky is not, as far as I know, optimistic about AI, but he shares the computer model with AI[1]), and the computer model was not invented by AI. If it was invented by anyone, it was invented by Alan Turing. Computer science is not the same thing as AI.

In fact, the idea of the mind as a sort of reckoning machine goes back to the seventeenth century.[2] In the early twentieth century two giants in logic—Kurt Gödel and Jacques Herbrand—first proposed the modern conception of computability (under the name "general recursiveness"[3]). Turing reformulated the Gödel-Herbrand notion of computability in terms that connect directly with digital computers (which were not yet invented, however!) and also suggested his

Hilary Putnam is Walter Beverly Pearson Professor of Modern Mathematics and Mathematical Logic in the department of philosophy at Harvard University.

abstract computers as a model for a mind.[4] Even if Turing's suggestion should prove wrong—even if it should prove in some way more empty than it seems—it would still have been a great contribution to thinking in the way past models of the mind have proved great contributions to thinking—great, even if not finally successful, attempts to understand understanding itself. But AI is not recursion theory, is not the theory of Turing machines, is not the philosophy of Alan Turing, but is something much more specific.

To get to AI, we first have to get to computers. The modern digital computer is a realization of the idea of a universal Turing machine in a particularly effective form—effective in terms of size, cost, speed, and so on. The construction and improvement of computers in terms of both software and hardware is a fact of life. But not everyone concerned with the design of either software or hardware is an AI researcher. However, some of what AI gets credit for—for example, the enormous improvement in the capacities of chess-playing computers—is as much or more due to discoveries of the inventors of hardware as it is to anything that might be called a discovery in AI.

Computer design is a branch of engineering (even when what is designed is software and not hardware), and AI is a subbranch of this branch of engineering. If this is worth saying, it is because AI has become notorious for making exaggerated claims—claims of being a fundamental discipline and even of being "epistemology." The aim of this branch of engineering is to develop software that will enable computers to simulate or duplicate the achievements of what we intuitively recognize as "intelligence."

I take it that this is a noncontroversial characterization of AI. The next statement I expect to be more controversial: AI has so far spun off a good deal that is of real interest to computer science in general, but nothing that sheds any real light on the mind (beyond whatever light may already have been shed by Turing's discussions). I don't propose to spend my pages defending this last claim (Joseph Weizenbaum has already done a good job along these lines[5]). But I will give a couple of illustrations of what I mean.

Many years ago I was at a symposium with one of the most "famous names" in AI. The famous name was being duly "modest" about the achievements of AI. He said offhandedly, "We haven't really achieved so much, but I will say that we now have *machines that understand children's stories.*" I remarked, "I know the program you refer to" (it was one of the earliest language-recognition programs). "What you didn't mention is that the program has to be *revised* for each new children's story." (That is, in case the point hasn't been grasped, the "program" was a program for answering questions about a specific children's story, not a program for understanding children's stories in general.) The famous name dropped the whole issue in a hurry.

Currently the most touted achievement of AI is "expert systems." But these systems (which are, at bottom, just high-speed data-base searchers) are not models for any interesting mental capacities.

Of course, the possibility remains that some idea dreamed up in an AI lab may in the future revolutionize our thinking about some aspect of mentation. (Parallel distributed processing is currently exciting interest as a possible model for at least some mental processes, for example. This is not surprising, however, since the model was suggested in the first place by the work of the neurologist D.O. Hebb.[6]) My point is not to predict the future but just to explain why I am inclined to ask, What's all the fuss about *now*? Why a whole issue of *Dædalus*? Why don't we wait until AI achieves something and *then* have an issue?

"IN PRINCIPLE"/"IN PRACTICE"

Perhaps the issue that interests people *is* whether we can model the mind or brain as a digital computer—in principle as opposed to right now—and perhaps AI gets involved because people do not sharply distinguish the in-principle question from the empirical question, Will AI succeed in so modeling the mind or brain? It may be useful to begin by seeing just how different the two questions are.

In one way the difference seems obvious: we are tempted to say that it might be possible in principle to model the mind or brain as a digital computer with appropriate software, but it might be too difficult in practice to write down the correct software. Or it just looks as if this difference is obvious. I want to say, Tread lightly; things are not so simple: in one sense, any physical system can be modeled as a computer.[7] The claim that the brain can be modeled as a computer is thus, in one way, trivial. Perhaps there is another more meaningful sense in which we can ask, Can the brain be modeled as a computer? At this point, however, all we can say is that the sense of the question has not been made clear.

But the feeling seems to be that not only is it possible in principle to model the mind or brain computationally, but there is a very good chance that we will be able to do it in practice, and philosophers (and defectors from AI like Weizenbaum) are seen as reactionaries who might talk us out of even trying something that promises to be a great intellectual and practical success. If this is how one thinks, then the gap between the two questions (and the vagueness of the in-principle question) may not seem very important in practice. Indeed, it may be of strategic benefit to confuse them.

The reasons for expecting us to succeed in practice are not clear to me, however.[8] If we are digital computers programmed by evolution, then it is important to know how to think about evolution. The great evolutionary biologist François Jacob once compared evolution to a tinker.[9] Evolution should not, Jacob wrote, be thought of as a designer who sits down and produces a lovely blueprint and then constructs organisms according to the blueprint. Evolution should rather be thought of as a tinker with a shop full of spare parts, interesting "junk," etc. Every so often the tinker gets an idea: "I wonder if it would work if I tried using this bicycle wheel in that doohickey?" Many of the tinker's bright ideas fail, but every so often one works. The result is organisms with many arbitrary features as well as serendipitous ones.

Now, imagine that the tinker becomes a programmer. Still thinking like a tinker, he develops "natural intelligence," not by writing a Grand Program and then building a device to realize it but by introducing one device or programming idea after another. (Religious people often reject such a view, for they feel that if it is right, then our nature and history is all "blind chance," but I have never been able to sympathize with this objection. Providence may work through what Kant called "the cunning of Nature.") The net result could be that natural intelligence is not the expression of some *one* program but the expression of billions of bits of "tinkering."

Something like this was, indeed, at one time discussed within the AI community itself. This community has wobbled back and forth between looking for a Master Program (ten or fifteen years ago there was a search for something called inductive logic) and accepting the notion that "artificial intelligence is one damned thing after another."

My point is that if AI is "one damned thing after another," the number of "damned things" the tinker may have thought of could be astronomical.[10] The upshot is pessimistic indeed: if there is no Master Program, then we may never get very far in terms of simulating human intelligence. (Of course, some areas that are relatively closed—for example, theorem proving in pure mathematics—might be amenable. Oddly enough, theorem proving has always been a rather underfunded part of AI research.)

A MASTER PROGRAM?

But why shouldn't there be a Master Program? In the case of deductive logic, we have discovered a set of rules that satisfactorily formalize valid inference. In the case of inductive logic, we have found no such rules, and it is worthwhile pausing to ask why.

In the first place, it is not clear just how large the scope of inductive logic is supposed to be. Some writers consider the "hypothetico-deductive method"—that is, the inference from the success of a theory's predictions to the acceptability of the theory—the most important part of inductive logic, while others regard it as already belonging to a different subject. Of course, if by "induction" we mean any method of valid inference that is not deductive, then the scope of the topic "inductive logic" will be enormous.

If the success of a large number (say, a thousand or ten thousand) of predictions that were not themselves consequences of auxiliary hypotheses alone (and that were unlikely in relation to what background knowledge gives us, Karl Popper would add[11]) always confirmed a theory, then at least the hypothetico-deductive inference would be easy to formalize. But problems arise at once. Some theories are accepted when the number of confirmed predictions is still very small. This was the case with the general theory of relativity, for example. To take care of such cases, we postulate that it is not only the number of confirmed predictions that matters but also the elegance or simplicity of the theory in question. Can such quasi-aesthetic notions as "elegance" and "simplicity" really be formalized? Formal measures have indeed been proposed, but it cannot be said that they shed any light on real-life scientific inference. Moreover, a confirmed theory sometimes fits badly with background knowledge; in some cases we conclude that the theory cannot be true, while in others we conclude that the background knowledge should be modified. Again, apart from imprecise talk about simplicity, it is hard to say what determines whether it is better in a particular case to preserve background knowledge or to modify it. And even a theory that leads to a vast number of successful predictions may not be accepted if someone points out that a much simpler theory would lead to those predictions as well.

In view of these difficulties, some students of inductive logic would confine the scope of the subject to simpler inferences, such as the inference from the statistics for a sample drawn from a population to the statistics for the entire population. When the population consists of objects that exist at different times, including future times, the present sample is never going to be a random selection from the whole population, however, so the key case is this: I have a sample that is a random selection from the members of a population who exist now (or worse, from the ones who exist here, on Earth, in the United States, in the particular place where I have been able to gather samples, or wherever). What can I conclude about the properties of future members of that population (and about the properties of members in other places)?

If the sample is a sample of uranium atoms, and the future members are in the near as opposed to the cosmological future, then we are prepared to believe that the future members will resemble present members, on the average. If the sample is a sample of people, and the future members of the population are not in the very near future, then we are less likely to make this assumption, at least if culturally variable traits are in question. Here we are guided by background knowledge, of course. This sort of example has suggested to some inquirers perhaps all there is to induction is the skillful use of background knowledge—we just "bootstrap" our way from what we know to additional knowledge. But then the cases in which we don't have much background knowledge, as well as the exceptional cases in which what we have to do is precisely question background knowledge, assume great importance; and here, as just remarked, no one has much to say beyond vague talk about simplicity.

The problem of induction is not by any means the only problem confronting anyone who seriously intends to simulate human intelligence. Induction—indeed, all cognition—presupposes the ability to recognize similarities among things; but similarities are by no means just constancies of the physical stimulus or patterns in the input to the sense organs. What makes knives similar, for example, is not that they all look alike (they don't), but that they are all manufactured to cut or stab (I neglect such cases as ceremonial knives here, of course). Thus, any system that can recognize knives as relevantly similar must be able to attribute *purposes* to agents. Humans have no difficulty in doing this. But it is not clear that we do this by unaided induction; we may well have a "hard-wired-in" ability to put ourselves in the shoes of other people that enables us to attribute to them any purposes we are capable of attributing to ourselves—an ability that Evolution the Tinker found it convenient to endow us with and one that helps us to know which of the infinitely many possible inductions we might consider is likely to be successful. Again, to recognize that a Chihuahua and a Great Dane are similar in the sense of belonging to the same species requires the ability to realize that, appearances notwithstanding,[12] Chihuahuas can impregnate Great Danes and produce fertile offspring. Thinking in terms of potential for mating and for reproduction is natural for us, but it need not be natural for an artificial intelligence—unless we deliberately simulate this human propensity when we construct the artificial intelligence. Such examples can be multiplied indefinitely.

Similarities expressed by adjectives and verbs rather than by nouns can be even more complex. A nonhuman intelligence might know what "white" is on a color chart, for example, without being able to see why pinkish gray humans are called white, and it might know what it is to open a door without being able to understand why we speak of opening a border or opening trade. There are many words (as Ludwig Wittgenstein pointed out[13]) that apply to things that have only a "family resemblance" to one another; there need not be one thing all *x*'s have in common. For example, we speak of the Canaanite tribal chiefs of the Old Testament as kings although their kingdoms were probably little more than villages, and we speak of George VI as a king, though he did not literally rule England; we even say that in some cases in history, kingship has not been hereditary. Similarly (in Wittgenstein's example), there is no property all games have in common that distinguishes them from all the activities that are not games.

The notional task of artificial intelligence is to simulate intelligence, not to duplicate it. So perhaps one might finesse the problems just mentioned by constructing a system that reasoned in an ideal language[14]—one in which words did not change their extensions in a context-dependent way (a sheet of typing paper might be "white$_1$" and a human being might be "white$_2$" in such a language, where "white$_1$" is color-chart white and "white$_2$" is pinkish gray). Perhaps all family-resemblance words would have to be barred from such a language. (How much of a vocabulary would be left?) But my list of difficulties is not yet finished.

Because the project of symbolic inductive logic appeared to run out of steam after Rudolf Carnap, the thinking among philosophers of science has, as I reported, run in the direction of talking about bootstrapping methods—methods that attribute a great deal to background knowledge. It is instructive to see why philosophers have taken this approach and also to realize how unsatisfactory it is if our aim is to simulate intelligence rather than to describe it.

One huge problem might be described as the existence of conflicting inductions. Here's an example from Nelson Goodman: as far as we know, no one who has ever entered Emerson Hall at Harvard University has been able to speak Inuit (Eskimo). This statement suggests the induction that if any person enters Emerson Hall, then he or she does not speak Inuit.[15] Let Ukuk be an Eskimo in Alaska who speaks Inuit. Shall I predict that if Ukuk enters Emerson Hall, Ukuk will no longer be able to speak Inuit? Obviously not, but what is wrong with this induction?

Goodman answers that what is wrong with the inference is that it conflicts with the "better entrenched," inductively supported law that people do not lose their ability to speak a language upon entering a new place. But how am I supposed to know that this law does have more confirming instances than the regularity that no one who enters Emerson Hall speaks Inuit? Through background knowledge again?

As a matter of fact, I don't believe that as a child I had any idea how often either of the conflicting regularities in the example (conflicting in that one of them must fail if Ukuk enters Emerson Hall) had been confirmed, but I would still have known enough not to make the silly induction that Ukuk would stop being able to speak Inuit if he entered a building (or a country) where no one had spoken Inuit. Again, it is not clear that the knowledge that one doesn't lose a language just like that is really the product of induction; perhaps this is something we have an innate propensity to believe. The question that won't go away is *how much what we call intelligence presupposes the rest of human nature.*

Moreover, if what matters really is "entrenchment" (that is, the number and variety of confirming instances), and if the information that the universal statement "One doesn't lose one's ability to speak a language upon entering a new place" is better entrenched than the universal statement "No one who enters Emerson Hall speaks Inuit" is part of my background knowledge, it isn't clear how that information got there. Perhaps the information is implicit in the way people speak about linguistic abilities; but then one is faced with the question of how one decodes the implicit information conveyed by the utterances one hears.

The problem of conflicting inductions is ubiquitous even if one restricts attention to the simplest inductive inferences. If the solution is really just to give the system more background knowledge, then what are the implications for artificial intelligence?

It is not easy to say, because artificial intelligence as we know it doesn't really try to simulate intelligence at all. Simulating intelligence is only its notional activity; its real activity is writing clever programs for a variety of tasks. But if artificial intelligence existed as a real, rather than notional, research activity, there would be two alternative strategies its practitioners could follow when faced with the problem of background knowledge:

1. They could accept the view of the philosophers of science I have described and simply try to program into a machine all the information a sophisticated human inductive judge has (including implicit information). At the least, this would require generations of researchers to formalize the information (probably it could not be done at all, because of the sheer quantity of information involved), and it is not clear that the result would be more than a gigantic expert system. No one would find this very exciting, and such an "intelligence" would in all likelihood be dreadfully unimaginative, unable to realize that in many cases it is precisely background knowledge that needs to be given up.

2. AI's practitioners could undertake the more exciting and ambitious task of constructing a device that could learn the background

knowledge by interacting with human beings, as a child learns a language and all the cultural information, explicit and implicit, that comes with learning a language by growing up in a human community.

THE NATURAL-LANGUAGE PROBLEM

The second alternative is certainly the project that deserves the name "artificial intelligence." But consider the problems: to figure out what is the information implicit in the things people say, the machine must simulate understanding a human language. Thus, the idea of sticking to an artificial ideal language and ignoring the complexities of natural language has to be abandoned if this strategy is adopted—abandoned because the cost is too high. Too much of the information the machine would need is retrievable only via natural-language processing.

But the natural-language problem presents many of the same difficulties all over again. Chomsky and his school believe that a "template" for natural language, including the "semantic," or conceptual, aspects, is innate—hard-wired-in by Evolution the Tinker.[16] Although this view is taken to extremes by Jerry Fodor, who holds that there is an innate language of thought with primitives adequate for the expression of all concepts that humans are able to learn to express in a natural language,[17] Chomsky himself has hesitated to go this far. What Chomsky seems committed to is the existence of a large number of innate conceptual abilities that give us a propensity to form certain concepts and not others. (In conversation, he has suggested that the difference between postulating innate concepts and postulating innate abilities is not important if the postulated abilities are sufficiently structured.) At the opposite extreme there is the view of classical behaviorism, which explains language learning as a special case of the application of general rules for acquiring "habits"—that is, as just one more bundle of inductions. (An in-between position is, of course, possible: Why should language learning not depend partly on special-purpose heuristics and partly on general learning strategies, both developed by evolution?)

Consider the view that language learning is not really learning but rather the maturation of an innate ability in a particular environment (somewhat like the acquisition of a birdcall by the young of a species of bird that has to hear the call from adult birds of the species to acquire it but that also has an innate propensity to acquire that sort of call). In its extreme form, this view leads to pessimism about the likelihood that the human use of natural language can be successfully simulated on a computer. This is why Chomsky is pessimistic about projects for natural-language computer processing, although he shares the computer model of the mind, or at least of the "language organ," with AI researchers. Notice that this pessimistic view about language learning parallels the pessimistic view that induction is not a single ability but rather a manifestation of a complex human nature whose computer simulation would require a vast system of subroutines—so vast that generations of researchers would be required to formalize even a small part of the system.

Similarly, the optimistic view that there is an algorithm of manageable size for inductive logic is paralleled by the optimistic view of language learning. This is the idea that there is a more or less topic-neutral heuristic for learning and that this heuristic suffices (without the aid of an unmanageably large stock of hard-wired-in background knowledge or topic-specific conceptual abilities) for learning one's natural language as well as for making inductive inferences. Perhaps the optimistic view is right, but I do not see anyone on the scene, in either artificial intelligence or inductive logic, who has any interesting ideas about how the topic-neutral learning strategy works. When someone does appear with such an idea, that will be the time for *Dædalus* to publish an issue on AI.

ENDNOTES

[1] Noam Chomsky, *Modular Approaches to the Study of the Mind* (San Diego, Calif.: San Diego State University Press, 1983).

[2] This is well described in Justin Webb's *Mechanism, Mentalism, and Metamathematics* (Dordrecht: Reidel, 1980).

[3] The Gödel-Herbrand conception of recursiveness was further developed by Stephen Kleene, Alonzo Church, Emil Post, and Alan Turing. The identification of recursiveness with effective computability was suggested (albeit obliquely) by Kurt Gödel in "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I." The German original of this was published in the *Monatshefte für Mathematik und Physik* 38 (1931):173–98; the English translation is in *The Undecidable: Basic Papers on Undecidable Propositions, Undecidable Problems, and Computable Functions,* ed. Martin Davis (Hewlett, N.Y.: Raven Press, 1965), 5–38. The idea was then explicitly put forward by Church in

his classic paper on the undecidability of arithmetic, "A Note on the Entscheidungsproblem," *Journal of Symbolic Logic* 1 (1) (March 1936):40–41; correction, ibid. (3) (September 1936):101–102; reprinted in Davis, *The Undecidable,* 110–15.

[4] Alan Turing and Michael Woodger, *The Automatic Computing Machine: Papers by Alan Turing and Michael Woodger* (Cambridge: MIT Press, 1985).

[5] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: Freeman, 1976).

[6] See David E. Rummelhart and James L. McClelland and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* vols. 1 and 2 (Cambridge: MIT Press, 1986); and D. O. Hebb, *Essay on Mind* (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980).

[7] More precisely, if we are interested in the behavior of a physical system that is finite in space and time and we wish to predict that behavior only up to some specified level of accuracy, then (assuming that the laws of motion are themselves continuous functions) it is trivial to show that a step function will give the prediction to the specified level of accuracy. If the possible values of the boundary parameters are restricted to a finite range, then a finite set of such step functions will give the behavior of the system under all possible conditions in the specified range to within the desired accuracy. But if that is the case, the behavior of the system is described by a recursive function and hence the system can be simulated by an automaton.

[8] In his reply to this paper (in this very issue), Daniel Dennett accuses me of offering an "a priori" argument that success is impossible. I have not changed the text of the paper at all in the light of his reply, and I invite the reader to observe that no such "a priori proof of impossibility" claim is advanced by me here or elsewhere! Although Dennett says that he is going to explain what AI has taught us about the mind, what he in fact does is to repeat the insults that AI researchers hurl at philosophers ("We are experimenters, and you are armchair thinkers!"). On other occasions, when Dennett is not talking like a spokesman for AI but doing what he does best, which is philosophy, he is, of course, well aware that I and, for that matter, other philosophers he respects are by no means engaged in a priori reasoning, and that the fact that we do not perform "experiments" does not mean that we are not engaged—as he is—in thinking about the real world in the light of the best knowledge available.

[9] François Jacob, "Evolution and Tinkering," *Science* 196 (1977):1161–66.

[10] That the number of times our design has been modified by evolution may be astronomical does not mean that the successful modifications are not (partially) hierarchically organized, nor does it mean that there are not a great many principles that explain the functioning together of the various components. To describe the alternative to the success of AI as "the mind as chaos," as Dennett does, is nonsense. If it turns out that the mind is chaos when modeled as a computer, that will only show that the computer formalism is not a perspicuous formalism for describing the brain, not that the brain is chaos.

[11] Karl Popper, *The Logic of Scientific Discovery* (London: Hutchinson, 1959).

[12] Note that if we had only appearances to go by, it would be quite natural to regard Great Danes and Chihuahuas as animals of different species!

[13] See Ludwig Wittgenstein, *Philosophical Investigations* (Oxford: Basil Blackwell, 1958), sec. 66–71.

[14] Note that this idea was one of the foundation stones of logical positivism. Although the positivists' goal was to reconstruct scientific reasoning rather than to mechanize it, they ran into every one of the problems mentioned here; in many ways the history of artificial intelligence is a repeat of the history of logical positivism (the second time perhaps as farce).

[15] Nelson Goodman, *Fact, Fiction, and Forecast,* 4th ed. (Cambridge: Harvard University Press, 1983).

[16] Chomsky speaks of "a subsystem [for language] which has a specific integrated character and which is in effect the genetic program for a specific organ" in the discussion with Seymour Papert, Jean Piaget, et al. reprinted in *Language and Learning,* ed. Massimo Piatelli (Cambridge: Harvard University Press, 1980). See also Noam Chomsky, *Language and Problems of Knowledge, The Managua Lectures* (Cambridge: MIT Press, 1987).

[17] Jerry A. Fodor, *The Language of Thought* (New York: Thomas Y. Crowell, 1975).