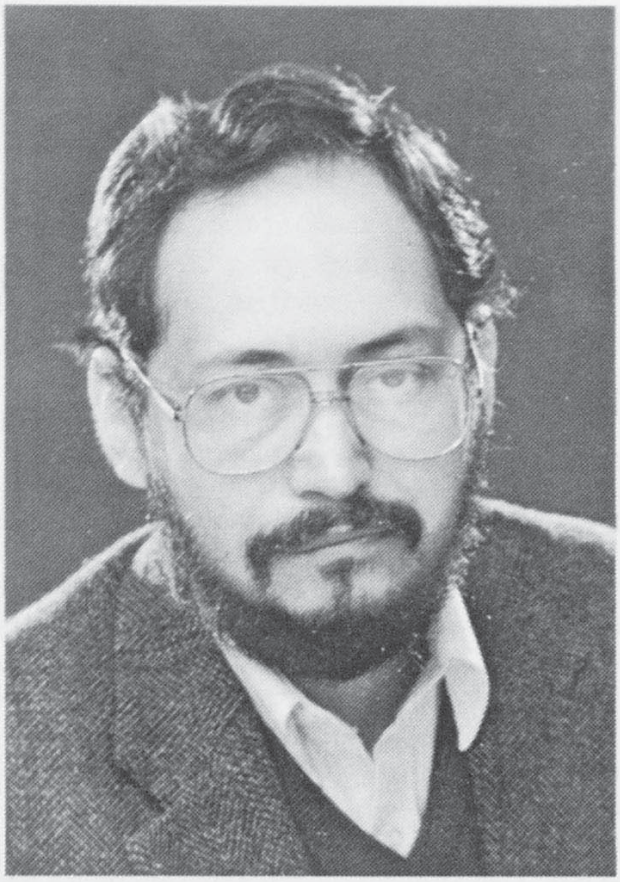

Gonzalo Muñevar

Philosophy and Exploration of the Solar System

Dr. Gonzalo Muñevar

The recent discovery of planets around other stars and of organic carbon in a Martian meteorite has renewed scientific interest in the search for extraterrestrial intelligence (SETI). I intend to examine one important argument against SETI that touches on some issues from the philosophy of science, particularly of AI and of biology.

Inspired by Copernicus, *proponents* of SETI assume the **principle of mediocrity**, which asserts that the sun is a typical star in having a planet like the Earth, propitious to the origin of life, that terrestrial life is typical in having produced intelligence, and that human intelligence is typical in giving rise to a technological civilization.[1] The *opponents* of SETI stretch this principle slightly to add that a technological civilization is typically expansionist. As a result they are able to produce a variety of "impossibility proofs" against the existence of extraterrestrial intelligence (ETI). Although the principle of mediocrity itself is in need of philosophical criticism, I will restrict my remarks to the main impossibility proof.

If ETIs do exist, Fermi once asked, why aren't they here? The argument is, briefly, that if the SETI proponents are right, there should be technological civilizations far older, and presumably far more advanced than ours; for many "typical" stars in our galaxy are billions of years older than the sun, and thus in some of their planets intelligence should have sprung long before it did on Earth. Now, just as we expanded from our beginnings in Africa, any civilization capable of space flight is bound to expand throughout the galaxy. Furthermore, this expansion would take place in a short amount of time (travelling at .1c, which is within human reach now, it would take only one million years to cross the galaxy). Thus the ETIs should be everywhere in the galaxy by now. But they clearly are not here; therefore there are no extraterrestrials in this galaxy.[2]

These opponents realize that interstellar travel may be too long and arduous for biological beings. They also realize that it may be unfeasible, even for advanced civilizations, to send "unmanned" probes to the — possibly — billions of planets in the galaxy. Their impossibility proof depends instead on a technology they believe is inevitable: self-reproducing machines.

John von Neumann supposedly proved already that a machine could be designed to make copies of itself.[3] A more advanced civilization would surely have discovered the equivalent of von Neumann's proof, and would be in a position to develop the appropriate technology. Indeed, some NASA scientists are investigating the possibility of using such machines to explore the galaxy.[4] We already have a mathematical proof that self- replicating machines can exist, all we need is the talent, effort, and money to create the technology. If this seems feasible for us, an advanced technological civilization surely would have no trouble building a couple. Once they arrive at their destinations, these machines would make copies of themselves, which would then move on to the nearest stars, and so on, setting in motion a geometric progression, until they overrun the entire galaxy. But we have no evidence of such machines here; therefore no advanced civilizations exist.

Von Neumann offered not one but five proofs. The first one, however, is the main basis on which all these promising speculations rest. Von Neumann knew that, through evolution, organisms produce others more complicated than themselves, and he wondered whether the opposite had to happen with machines. So he set out to determine whether it was possible to program a computer to make a copy of itself. He imagined a robot floating in a vat full of robot parts. The robot could be programmed

to pick up a part and identify it. Then the robot, which had a blueprint of itself, would look for the connecting parts, and would then begin putting another robot together the way a child puts together a Mechano set. Surely, we can also program a computer to do this (it is far more complicated, but possible). Once the second robot was assembled, the first would pass on to it the self-replicating program (or set of programs, rather). By breaking down the task of self-replication into small, manageable tasks, von Neumann thought, an automaton could copy itself. This result led him to remark that there were two kinds of automata: artificial automata, such as computers, and natural automata, such as people and cats.

Thus the implications of this very simple conceptual (rather than mathematical) proof go well beyond the concerns of technology and exploration — they affect also our notion of life. But let us see what happens when we send one of von Neumann's self-replicating automata (SRAs) into the galaxy.

The first thing that comes to mind is that when an SRA gets to another world it is not going to land in a vat full of parts. It will have to build factories to build the parts from the raw materials that it will mine. But the factories are themselves made of parts, so it will have to build other factories to build the parts to build the factories.... This is called the "closure problem" by those in the field. There is no need to fear an infinite regress, however, for we know that the closure problem can be solved — even if we still have no idea how to program a machine to solve it. And we know that it can be solved because our technological civilization solves it: we do send rockets to other worlds.

We must realize, however, that such an SRA will be an extremely complex machine, both in its computer programs and in its physical realization. Indeed, an SRA will be the equivalent of a technological civilization (including the starship by which it moves about). Whether we can write a program that complex, and whether we can assemble and then make to work a machine that sophisticated is open to question. But let me assume for the sake of argument that we can.

Let us imagine what would happen when one of these extremely complex SRAs arrives at a planetary system. Now, we do not yet know what other planetary systems look like, but some theories suggest that they would be collections of small rocky planets and gas giants. Let us suppose now that an SRA comes into *our* solar system. Jupiter and Saturn would not be good places to land (even figuratively) because it is unlikely that the SRA can fashion the needed parts and factories out of the hydrogen, helium, and the trace gases that can be found in their atmospheres. The moons of the gas giants are not that much better, for surely a machine equivalent to a technological civilization may be presumed to need a variety of materials, including metals, for the task of self-replication. Unfortunately the low density of these moons (less than 2.0 g/cm^3) suggests that they would not be good places to search for the necessary raw materials.

Nevertheless we know that in rocky planets like Earth an SRA can find practically everything it needs (or so we hope). But even on rocky planets the SRA's problems are far from over. Small differences between the planets in astrophysical terms may lead to significant differences in density and chemical composition of the atmosphere. These significant differences will in turn make it necessary to adopt different strategies for mining and manufacturing. For example, on Earth the best way to treat some particular ore may be to throw it into a pot of boiling water. In Mars the water would evaporate before the ore is settled in. In Venus the pot itself might melt.

This is by no means a small problem. No matter how similar planetary systems may be to one another, we should still expect at least some small astrophysical differences between their rocky planets. Thus the possible combinations of factors that may affect

mining and manufacturing may be practically infinite. Therefore the already extremely complex SRA would need, in addition, some general purpose programs so it can begin the task of making the parts for its progeny. But no one knows how to write such a general purpose program, and there are reasons for thinking that they cannot be written.[5] The biggest stumbling block to artificial intelligence has been precisely the inability to write programs that exhibit a flexible response to run-of-the-mill environments, let alone to the incredible variety demanded of SRAs. Nor is there any assurance that this problem can be overcome in the foreseeable future. (Connectionist approaches, which present a significant alternative to von Neumann's view, are better able to handle context, but this situation would still present too tall an order for them.)[6]

But if rocky planets are too heterogeneous for the SRAs' needs, we may still find a homogeneous environment where they can get all the raw materials in question: the asteroid belt. We may be stretching our luck by supposing that all planetary systems (or even most of them) have asteroid belts, but let that pass.

An SRA could move from asteroid to asteroid picking up metal ores here, carbon compounds there, mining and processing them all in a rather stable environment (the cold vacuum of interplanetary space — although the exact location of the asteroid belt, and the strength of the stellar wind in that system may again provide too much variety). After granting all this, we are now able to deal with the fundamental problem of SRAs. As von Neumann himself pointed out, the more complex a computer program is, the more likely it is to have errors. But the SRAs would be far more complex than anything we have ever imagined programming and building. These errors, furthermore, involve principally the task of self- replication. I am talking not only about bugs in the gigantic program but about errors of execution in manufacturing and assembling the many components (an alloy that is not quite up to strength, a tooth in a gear that is slightly short and with a bit of wear will no longer catch another as it must).

Neither quality control nor error-detecting programs will solve this problem, for it takes only a small percentage of error to bring the task of self-replication to a halt. Let me explain the difficulty by means of an illustration. The SRA is already saddled with a computer program so complex that it seems difficult to imagine that we can debug it completely. But now we must add to it a program that must equip the machine with ways of checking the complete specifications for all parts, all fittings, and all functions. Nonetheless even this added complexity does not solve the problem. For a program that can foresee all the possible ways in which something can go wrong (a piece of dust, a screw partly loose) begins to look like a general purpose program (that is one reason why astronauts are so useful in space). In a machine that complex, engaged in the extraordinarily complex task of producing another SRA, things can go wrong in more ways than we can imagine. A program that must deal with so many unknown contingencies is, again, a program that can deal with an open-ended environment. And that is where SRAs come to grief, once again.

But isn't it the case that living things, which are very complex in their own right, also make errors in the copying of the information used in replication? So why is it that error *must* bring machine replication to a halt when it does not do so in the replication of living things? The answer is as follows. In living things "errors" (mutations, recombinations) do serve to provide genetic variation in a population. Actually, many mutations are supposed to be deadly, but in some cases the genetic variation allows the population to adapt to changes in the environment; that is, as the environment changes, some of the members of the population may take advantage of past "errors," and the population lives on. This is how error can be adaptive for living things.

In the case of the SRAs, however, we must remember that the asteroid environment was chosen precisely because it would not change from one system to another. In that unchanging environment there is no advantage in error. In SRAs the bulk of the errors that concern us here are precisely in the reproductive part, and thus they are maladaptive. When all is said and done, it seems that an SRA technology is not really even a gleam in a scientist's eye.

Nevertheless many would think that we can actually point to examples of self-replicating machines: trees, cats, humans. These people already assume von Neumann's conclusion, that there are two kinds of automata — natural and artificial. I suspect that they find that assumption reasonable because they believe that the genetic code is the equivalent of a computer program, and thus they conclude that living things are just the realizations, or executions, of their particular programs. This view is no longer popular amongst biologists, but we still need to see why it fails to help the case of the SRAs.

First of all, if we are to use analogies, we should stress the following: in the case of the SRAs, the machine must make a copy of itself and then pass on the program (even if it passes on the program to a unit that is not yet completed, the point is that making the copy and passing on the program are separate, largely independent tasks); in the case of living things, however, it seems more proper to say that they pass on the program first, and then, as result of that, the copy is made. This is not a small difference, for in living things relatively simple accomplishments (e.g., a fertilized egg) can produce very complex organisms (the egg is extremely complex chemically, but simple relative to, say, the human being that will result from it).

This result leads to a second, and more important point. To picture the genetic code as a computer program is just to engage in metaphor, and the metaphor is highly misleading. The "instructions" of the DNA produce the expected results (e.g., proteins, cells, tissues, organs, behavior) only because at every level they can be expressed in appropriate environments, indeed it is often the appropriate environment that will trigger the next stage in embryonic development. In a human, for example, at a certain time in the life of the embryo the normal development requires a certain concentration of sodium, and after the human is born, the attention paid to him is not only necessary for his survival but provides the stimulation needed for the central nervous system to grow further.

In other words, the "instructions" of the DNA do not have meaning by themselves. This issue is similar to that of the meaning of words in the philosophy of language. It used to be thought that words had intrinsic meaning, but it is generally accepted now that the meaning of a word depends just as much on the context in which it is uttered (the meaning is given by the interaction of word and context, where the context may include a large variety of factors, including the relationship of the word to other elements of the sentence, the manner of its utterance, the social conditions that the speaker and the listeners take themselves to be in, etc.). In embryonic development, the "program" makes sense (has any meaning) only in that the maturing organisms interacts with a sequence of appropriate environments. Those environments provide the biological contexts in which the "instructions" of the genetic code are instructions at all.

Now, that sequence is itself the result of natural history, that is, of a long series of interactions between the ancestors of that organism and the environments in which they evolved. There is a clear sense, then, in which a living being comes into a world that is already made for it. The world of the SRA, on the other hand, must be largely described in its program from the beginning. The meaning of the program must be made explicit beforehand. To illustrate this point, let us consider the development of a nervous system.

In dissecting an animal we may find that its nerve cells always exhibit a certain pattern, and may thus imagine that pattern is contained in a blueprint in the DNA. Nevertheless, as the nerve cells grow through, say, a muscle tissue, they need not be guided by any such blueprint. They may simply have "instructions" to grow in the general direction of a chemical marker, until they make contact with a membrane, which will turn the "instructions" off. But the developing muscle cells will then constrain the manner in which the nerve cells grow (the nerve cells will have to grow around them, for example). The final pattern is the result of such contingencies, and there is no need for any blueprint whatsoever.

In living things the burden of development is largely assumed by natural history, in SRAs it mostly falls on the programmer's shoulders. That is why the first is manageable and the second is not. This is not to say that artificial life is impossible. As far as I can tell, there are no objections in principle against it. But it would be *life* nonetheless. Insofar as there is design in it, that design is grafted on to the knowledge we have of natural history, to take advantage of prior interactions with environments or sequences of environments. (I am not referring here to the computer field of "artificial life," based on Von Neumann's other proofs, which has conceptual problems of its own.)

I conclude that there is no good reason for thinking that a technology of self-reproducing machines is possible, much less practically inevitable. The impossibility proof fails.

Evergreen State College

NOTES

1 The classic expositor was Carl Sagan, see for example his (ed.) *Communication with Extraterrestrial Intelligence*, MIT Press, 1973.

2 Authors more contemporary than Fermi have developed the argument criticized here. See for example, Frank J. Tipler, "Extraterrestrial Beings do not Exist," *Physics Today*, April 1981, pp. 9-38. For recent commentary (and the reference to Fermi) see Paul Davies, *Are We Alone? Philosophical Implications of the Discovery of Extraterrestrial Life*, Basic Books, 1995, based on his series of lectures at the University of Milan in 1993.

3 John von Neumann, *Theory of Self-Reproducing Automata*, A.W. Burks, ed., University of Illinois Press, 1966.

4 *Advanced Automation for Space Missions*, NASA Conference Publication 2255, 1982.

5 The classic critique is Richard Dreyfus, *What Computers Can't Do*, Harper & Row, 1972.

6 Even in optimistic treatments such as Paul M. Churchland's, *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, MIT Press, 1989. I find Churchland's view very plausible, and in a sense this paper supports it by undermining the view of mind (and body) put forward by von Neumann.