# 16 • Learning from functionalism: prospects for strong artificial life

ELLIOTT SOBER

## TWO USES FOR COMPUTERS

There are two quite different roles that computers might play in biological theorizing. Mathematical models of biological processes are often analytically intractable. When this is so, computers can be used to get a feel for the model's dynamics. You plug in a variety of initial condition values and allow the rules of transition to apply themselves (often iteratively); then you see what the outputs are.

Computers are used here as aids to the theorist. They are like pencil and paper or a sliderule. They help you think. The models being investigated are about life. But there is no need to view the computers that help you investigate these models as alive themselves. Computers can be applied to calculate what will happen when a bridge is stressed, but the computer is not itself a bridge.

Population geneticists have used computers in this way since the 1960s. Many participants in the Artificial Life research program are doing the same thing. I see nothing controversial about this use of computers. By their fruits shall ye know them. This part of the AL research program will stand or fall with the interest of the models investigated. When it is obvious beforehand what the model's dynamics will be, the results provided by computer simulation will be somewhat uninteresting. When the model is very unrealistic, computer investigation of its properties may also fail to be interesting. However, when a model is realistic enough and the results of computer simulation are surprising enough, no one can deny the pay-off.

I just mentioned that computers may help us understand bridges, even though a computer is not a bridge. However, the second part of the artificial life research program is interested in the idea that computers are instances of biological processes. Here the computer is said to be alive, or to exemplify various properties that we think of as characteristic of life.

This second aspect of the artificial life research program needs to be clearly separated from the first. It is relatively uncontroversial that computers can be tools for investigating life; in contrast, it is rather controversial to suggest that computers are or can be alive. Neither of these ideas entails the other; they are distinct.

Shifting now from AL to AI, again we can discern two possible uses of computers. First, there is the use of computers as tools for investigating psychological models that are too mathematically complicated to be analytically solved. This is the idea of computers as tools for understanding the mind. Second and separate, there is the idea that computers are minds. This latter idea, roughly, is what usually goes by the name of strong AI.

Strong AI has attracted a great deal of attention, both in the form of advocacy and in the form of attack. The idea that computers, like paper and pencil, are tools for understanding psychological models has not been criticized much, nor should it have been. Here, as in the case of AL, by their fruits shall ye know them.

So in both AI and AL, the idea that computers are tools for investigating a theory is quite different from the idea that computers are part of the subject-matter of the theory. I'll have little more to say about the tool idea, although I'll occasionally harp on the importance of not confusing it with the subject-matter idea. Table 16.1 depicts the parallelism between AI and AL; it also should help keep the tool idea and the subject matter idea properly separated: Of course, "strong" does not mean plausible or well defended; "strong" means daring and "weak" means modest.

*The Nature of Life*, ed. M. A. Bedau and C. E. Cleland. Published by Cambridge University Press. 

Table 16.1. *Parallelism between AI and AL*

| | Psychology | Biology |
|---|---|---|
| Computers are tools for investigating | weak AI | weak AL |
| Computers are part of the subject matter of | strong AI | strong AL |

Where did the idea that computers could be part of the subject matter of psychological and biological theories come from? Recent philosophy has discussed the psychological issues a great deal, the biological problem almost not at all. Is it possible to build a case for living computers that parallels the arguments for thinking computers? How far the analogy can be pushed is what I want to determine.

## THE PROBLEM OF MIND AND THE PROBLEM OF LIFE

The mind/body problem has witnessed a succession of vanguard theories. In the early 1950s, Ryle[12] and Wittgenstein[18] advocated forms of logical behaviorism. This is the idea that the meaning of mentalistic terms can be specified purely in terms of behavior. Ryle attacked the Myth of the Ghost in the Machine, which included the idea that mental states are inner causes of outward behavior.

In the mid 1950s to mid 1960s, behaviorism itself came under attack from mind/brain identity theorists. Australian materialists like Place[10] and Smart[14] maintained that mental states are inner causes. But more than that, they argued that mental properties would turn out to be identical with physical properties. Whereas logical behaviorists usually argued for their view in an *a priori* fashion, identity theorists said they were formulating an empirical thesis that would be borne out by the future development of science.

The identity theory can be divided into two claims. They claimed that each mental object is a physical thing. They also claimed that each mental property is a physical property. In the first category fall such items as minds, beliefs, memory traces, and afterimages. The second category includes believing that snow is white and feeling pain. This division may seem a bit artificial—why bother to make separate claims about my belief that snow is white and the property of believing that snow is white? In a moment, the point of this division will become plain.

Beginning in the mid 1960s the identity theory was challenged by a view that philosophers called functionalism. Hilary Putnam,[11] Jerry Fodor,[6] and Daniel Dennett[3] argued that psychological properties are multiply realizable. If this is correct, then the identity theory must be rejected.

To understand what multiple realizability means, it is useful to consider an analogy. Consider mousetraps. Each of them is a physical object. Some are made of wire, wood, and cheese. Others are made of plastic and poison. Still others are constituted by bunches of philosophers scurrying around the room armed with inverted wastepaper baskets.

What do all these mousetraps have in common? Well, they are all made of matter. But more specifically, what properties do they share that are unique to them? If mousetraps are multiply realizable, then there is no physical property that all mousetraps, and only mousetraps, possess.

Each mousetrap is a physical thing, but the property of being a mousetrap is not a physical property. Here I am putting to work the distinction between object and property that I mentioned before.

Just as there are many physical ways to build a mousetrap, so, functionalists claimed, there are many physical ways to build a mind. Ours happens to be made of DNA and neurons. But perhaps computers could have minds. And perhaps there could be organisms in other species or in other galaxies that have minds, but whose physical organization is quite different from the one we exemplify. Each mind is a physical thing, but the property of having a mind is not a physical property.

Dualism is a theory that I have not mentioned. It claims that minds are made of an immaterial substance. Identity theorists reject dualism. So do functionalists. The relationship between these three theories can be represented by saying how each theory answers a pair of questions (see Table 16.2).

I mentioned earlier that identity theorists thought of themselves as advancing an empirical thesis about the nature of the mind. What about functionalism? Is it an empirical claim? Here one finds a division between two styles of argument. Sometimes functionalists

Table 16.2. *Three theories of mind*

| | | Are mental __ physical? | |
|---|---|---|---|
| | Dualism | Functionalism | Identity theory |
| Objects | No | Yes | Yes |
| Properties | No | No | Yes |

appear to think that the meaning of mentalistic terminology guarantees that the identity theory must be false. This *a priori* tendency within functionalism notwithstanding, I prefer the version of that theory that advances an empirical thesis. It is an empirical question how many different physical ways a thinking thing can be built. If the number is enormously large, then functionalism's critique of the identity theory will turn out to be right. If the number is very small (or even one), then the identity theory will be correct. Perhaps the design constraints dictated by psychology are satisfiable only within a very narrow range of physical systems; perhaps the constraints are so demanding that this range is reduced to a single type of physical system. This idea cannot be dismissed out of hand.

If philosophers during the same period of time had been as interested in the problem of life as they were in the problem of mind, they might have formulated biological analogs of the identity theory and functionalism. However, a biological analog of mind/body dualism does not have to be invented—it existed in the form of vitalism. Dualists claim that beings with minds possess an immaterial ingredient. Vitalists claim that living things differ from inanimate objects because the former contain an immaterial substance—an *élan vital.*

Had the problem of life recapitulated the problem of mind, the triumphs of molecular biology might then have been interpreted as evidence for an identity theory, according to which each biological property is identical with some physical property. Finally, the progression might have been completed with an analog of functionalism. Although each living thing is a material object, biological properties cannot be identified with physical ones.

Actually, this functionalist idea has been espoused by biologists, although not in the context of trying to recapitulate the structure of the mind/body problem. Thus, Fisher[5] says that "fitness, although measured by a uniform method, is qualitatively different for every different organism." Recent philosophers of biology have made the same point by arguing that an organism's fitness is the upshot of its physical properties even though fitness is not itself a physical property. What do a fit cockroach and a fit zebra have in common? Not any physical property, any more than a wood and wire mousetrap must have something physical in common with a human mouse catcher. Fitness is multiply realizable.

There are many other biologically interesting properties and processes that appear to have the same characteristic. Many of them involve abstracting away from physical details. For example, consider Lewontin's[7] characterization of what it takes for a set of objects to evolve by natural selection. A necessary and sufficient condition is heritable variation in fitness. The objects must vary in their capacity to stay alive and to have offspring. If an object and its offspring resemble each other, the system will evolve, with fitter characteristics increasing in frequency and less-fit traits declining.

This abstract skeleton leaves open what the objects are that participate in a selection process. Darwin thought of them as organisms within a single population. Group selectionists have thought of the objects as groups or species or communities. The objects may also be gametes or strands of DNA, as in the phenomena of meiotic drive and junk DNA.

Outside the biological hierarchy, it is quite possible that cultural objects should change in frequency because they display heritable variation in fitness. If some ideas are more contagious than others, they may spread through the population of thinkers. Evolutionary models of science exploit this idea. Another example is the economic theory of the firm; this describes businesses as prospering or going bankrupt according to their efficiency.

Other examples of biological properties that are multiply realizable are not far to seek. Predatory/prey theories, for example, abstract away from the physical details that distinguish lions and antelopes from spiders and flies.

I mentioned before that functionalism in the philosophy of mind is best seen as an empirical thesis about the degree to which the psychological characteristics of a system constrain the system's physical realization. The same holds for the analog of functionalism as a thesis about biological properties. It may be

somewhat obvious that some biological properties—like the property of being a predator—place relatively few constraints on the physical characteristics a system must possess. But for others, it may be much less obvious.

Consider, for example, the fact that DNA and RNA are structures by which organisms transmit characteristics from parents to offspring. Let us call them hereditary mechanisms. It is a substantive question of biology and chemistry whether other molecules could play the role of a hereditary mechanism. Perhaps other physical mechanisms could easily do the trick; perhaps not. This cannot be judged *a priori,* but requires a substantive scientific argument.

## WHAT ARE PSYCHOLOGICAL PROPERTIES, IF THEY ARE NOT PHYSICAL?

In discussing functionalism's criticism of the mind/brain identity theory, I mainly emphasized functionalism's negative thesis. This is a claim about what psychological states are not; they aren't physical. But this leaves functionalism's positive proposal unstated. If psychological properties aren't physical, what are they, then?

Functionalists have constructed a variety of answers to this question. One prominent idea is that psychological states are computational and representational. Of course, the interest and plausibility of this thesis depends on what "computational" and "representational" are said to mean. If a functionalist theory entails that desk calculators and photoelectric eyes have beliefs about the world, it presumably has given too permissive an interpretation of these concepts. On the other hand, if functionalism's critique of the identity theory is right, then we must not demand that a system be physically just like us for it to have a psychology. In other words, the problem has been to construct a positive theory that avoids, as Ned Block[1] once put it, both chauvinism and liberalism. Chauvinistic theories are too narrow, while liberal theories are too broad, in their proposals for how the domain of psychology is to be characterized.

## BEHAVIORISM AND THE TURING TEST

Functionalists claim that psychological theories can be formulated by abstracting away from the physical details that distinguish one thinking system from another. The question is: How much abstracting should one indulge in?

One extreme proposal is that a system has a mind, no matter what is going on inside it, if its behavior is indistinguishable from some other system that obviously has a mind. This is basically the idea behind the Turing Test. Human beings have minds, so a machine does too, if its behavior is indistinguishable from human behavior. In elaborating this idea, Turing was careful that "irrelevant" cues not provide a tip off. Computers don't look like people, but Turing judged this fact to be irrelevant to the question of whether they think. To control for this distracting detail, Turing demands that the machine be placed behind a screen and its behavior standardized. The behavior is to take the form of printed messages on a tape.

Besides intentionally ignoring the fact that computers don't look like people, this procedure also assumes that thinking is quite separate from doing. If intelligence requires manipulating physical objects in the environment, then the notion of behavior deployed in the Turing Test will be too meager. Turing's idea is that intelligence is a property of pure cogitation, so to speak. Behavior limited to verbal communication is enough.

So the Turing Test represents one possible solution to the functionalist's problem. Not only does thought not require a physical structure like the one our brains possess. In addition, there is no independently specifiable internal constraint of any kind. The only requirement is an external one, specified by the imitation game.

Most functionalists regard this test as too crude. Unfortunately, it seems vulnerable to both type-1 and type-2 errors. That is, a thing that doesn't think can be mistakenly judged to have a mind and a thinking thing can be judged to lack a mind by this procedure (see Table 16.3).

In discussions of AI, there has been little attention to type-2 errors. Yet, it seems clear that human beings with minds can imitate the behavior of mindless computers and so fool the interrogator into thinking that they lack minds. Of perhaps more serious concern is the possibility that a machine might have a mind, but have beliefs and desires so different from those of any human being that interrogators would quickly realize that they were not talking to a human being. The machine would flunk the Turing Test, because it cannot imitate human response patterns; it is another

Table 16.3. *Errors in the Turing Test*

| | The subject thinks | The subject does not think |
|---|---|---|
| The subject passes | ok | type-1 |
| The subject does not pass | type-2 | ok |

matter to conclude that the machine, therefore, does not have a mind at all.

The possibility of type-2 error, though real, has not been the focus of attention. Rather, in order to overcome the presumption that machines can't think, researchers in AI have been concerned to construct devices that pass the Turing Test. The question this raises is whether the test is vulnerable to type-1 errors.

One example that displays the possibility of type-1 error is due to Ned Block.[2] Suppose we could write down a tree structure in which every possible conversation that is 5 hours or less in duration is mapped out. We might trim this tree by only recording what an "intelligent" respondent might say to an interrogator, leaving open whether the interrogator is "intelligent" or not. This structure would be enormously large, larger than any current computer would be able to store. But let's ignore that limitation and suppose we put this tree structure into a computer.

By following this tree structure, the machine would interact with its interrogator in a way indistinguishable from the way in which an intelligent human being would do so. Yet the fact that the machine simply makes its way through this simple tree structure strongly suggests that the machine has no mental states at all.

One might object that the machine could be made to fail the Turing Test if the conversation pressed on beyond five hours. This is right, but now suppose that the tree is augmented in size, so that it encompasses all sensible conversations that are ten hours or less in duration. In principle, the time limit on the tree might be set at any finite size—four score and ten years if you like.

Block draws the moral that thinking is not fully captured by the Turing Test. What is wrong with this branching structure is not the behaviors it produces but how it produces them. Intelligence is not just the ability to answer questions in a way indistinguishable from that of an intelligent person. To call a behavior "intelligent" is to comment on how it is produced. Block concludes, rightly I think, that the Turing Test is overly behavioristic.

In saying this, I am not denying that the Turing Test is useful. Obviously, behavioral evidence can be telling. If one wants to know where the weaknesses are in a simulation, one might try to discover where the outputs mimic and where they do not. But it is one thing for the Turing Test to provide fallible evidence about intelligence, something quite different for the test to define what it is to have a mind.

A large measure of what is right about Searle's[13] much-discussed paper "Minds, Brains, and Programs" reduces to this very point. In Searle's Chinese room example, someone who speaks no Chinese is placed in a room, equipped with a manual, each of whose entries maps a story in Chinese (S) and a question in Chinese about that story (Q) onto an answer in Chinese to that story (A). That is, the man in the room has a set of rules, each with the form

$$S + Q \rightarrow A.$$

Chinese stories are sent into the room along with questions about those stories, also written in Chinese. The person in the room finds the $S + Q$ pairing on his list, writes down the answer onto which the pair is mapped, and sends that message out of the room. The input/output behavior of the room is precisely what one would expect of someone who understands Chinese stories and wishes to provide intelligent answers (in Chinese) to questions about them. Although the system will pass the Turing Test, Searle concludes that the system that executes this behavior understands nothing of Chinese.

Searle does not specify exactly how the person in the room takes an input and produces an output. The details of the answer manual are left rather vague. But Block's example suggests that this makes all the difference. If the program is just a brute-force pairing of stories and questions on the one hand and answers on the other, there is little inclination to think that executing the program has anything to do with understanding Chinese. But if the manual more closely approximates what Chinese speakers do when they answer questions about stories, our verdict might change. Understanding isn't definable in terms of the ability to answer questions; in addition, how one obtains the answers must be taken into account.

Searle considers one elaboration of this suggestion in the section of his paper called "The Brain Simulator Reply (Berkeley and MIT)."[13] Suppose we simulate "the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories in Chinese and gives answers to them." A system constructed in this way would not just duplicate the stimulus/response pairings exemplified by someone who understands Chinese; in addition, such a system would closely replicate the internal processes mediating the Chinese speaker's input/output connections.

I think that Searle's reply to this objection is question-begging. He says "the problem with the brain simulator is that it is simulating the wrong things about the brain. As long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states."

Although Searle does not have much of an argument here, it is important to recognize that the denial of his thesis is far from trivial. It is a conjecture that might or might not be right that the on/off states of a neuron, plus its network of connections with other neurons, exhaust what is relevant about neurons that allows them to form intentional systems. This is a meager list of neuronal properties, and so the conjecture that it suffices for some psychological characteristic is a very strong one. Neurons have plenty of other characteristics; the claim that they are all irrelevant as far as psychology goes may or may not be true.

There is another way in which Searle's argument recognizes something true and important, but, I think, misinterprets it. Intentionality crucially involves the relationship of "aboutness." Beliefs and desires are about things in the world outside the mind. How does the state of an organism end up being about one object, rather than about another? What explains why some states have intentionality whereas others do not? One plausible philosophical proposal is that intentionality involves a causal connection between the world and the organism. Crudely put, the reason my term "cat" refers to cats is that real cats are related to my use of that term in some specific causal way. Working out what the causal path must be has been a difficult task for the causal theory of reference. But leaving that issue aside, the claim that some sort of causal relation is necessary, though perhaps not sufficient, for at least some of the concepts we possess, has some plausibility.

If this world/mind relationship is crucial for intentionality, then it is clear why the formal manipulation of symbols can't, by itself, suffice for intentionality. Such formal manipulation is purely internal to the system, but part of what makes a system have intentionality involves how the system and its states are related to the world external to the system.

Although conceding this point may conflict with some pronouncements by exponents of strong AI, it does not show that a thinking thing must be made of neuronal material. Consistent with the idea that intentionality requires a specific causal connection with the external world is the possibility that a silicon-chip computer could be placed in an environment and acquire intentional states by way of its interactions with the environment.

It is unclear what the acquisition process must be like, if the system is to end up with intentionality. But this lack of clarity is not specific to the question of whether computers could think; it also applies to the nature of human intentionality. Suppose that human beings acquire concepts like *cat* and *house;* that is, suppose these concepts are not innate. Suppose further that people normally acquire these concepts by causally interacting with real cats and houses. The question I wish to raise is whether human beings could acquire these concepts by having a neural implant performed at birth that suitably rewired their brains. A person with an implant would grow up and feel about the world the way any of us do who acquired the concepts by more normal means. If artificial interventions in human brains can endow various states with intentionality, it is hard to see why artificial interventions into silicon computers can't also do the trick.

I have tried to extract two lessons from Searle's argument. First, there is the idea that the Turing Test is overly behavioristic. The ability to mimic intelligent behavior is not sufficient for having intelligence. Second, there is the idea that intentionality—aboutness—may involve a world/mind relationship of some specifiable sort. If this is right, then the fact that a machine (or brain) executes some particular program is not sufficient for it to have intentionality. Neither of these conclusions shows that silicon computers couldn't have minds.

What significance do these points have for the artificial life research program? Can the biological properties of an organism be defined purely in terms of environment/behavior pairings? If a biological property has this characteristic, the Turing Test will work better for it than the test works for the psychological property that Turing intended to describe. Let's consider photosynthesis as an example of a biological process. Arguably, a system engages in photosynthesis if it can harness the Sun's energy to convert water and $CO_2$ into simple organic compounds (principally $CH_2O$). Plants (usually) do this in their chloroplasts, but this is just one way to do the trick. If this is right, either for photosynthesis or for other biological properties, then the first lesson I drew about the Turing Test in the context of philosophy of mind may actually provide a disanalogy between AI and AL. Behaviorism is a mistake in psychology, but it may be the right view to take about many biological properties.

The second lesson I extracted from Searle's argument concerned intentionality as a relationship between a mental state and something outside itself. It is the relation of aboutness. I argued that the execution of a program cannot be the whole answer to the question of what intentionality is.

Many biological properties and processes involve relationships between an organism (or part of an organism) and something outside itself. An organism reproduces when it makes a baby. A plant photosynthesizes when it is related to a light source in an appropriate way. A predator eats other organisms. Although a computer might replicate aspects of such processes that occur inside the system of interest, computers will not actually reproduce or photosynthesize or eat unless they are related to things outside themselves in the right ways. These processes involve actions—interactions with the environment; the computations that go on inside the skin are only part of the story. Here, then, is an analogy between AI and AL.

## THE DANGER OF GOING TOO FAR

Functionalism says that psychological and biological properties can be abstracted from the physical details concerning how those properties are realized. The main problem for functionalism is to say how much abstraction is permissible. A persistent danger for functionalist theories is that they err on the side of being too liberal. This danger is especially pressing when the mathematical structure of a process is confused with its empirical content. This confusion can lead one to say that a system has a mind or is alive (or has some more specific constellation of psychological or biological properties) when it does not.

A simple example of how this fallacy proceeds may be instructive. Consider the Hardy–Weinberg Law in population genetics. It says what frequencies the diploid genotypes at a locus will exhibit, when there is random mating, equal numbers of males and females, and no selection or mutation. It is, so to speak, a "zero-force law"—it describes what happens in a population if no evolutionary forces are at work (see Sober[16]). If $p$ is the frequency of the $A$ allele and $q$ is the frequency of $a$ (where $p + q = 1$), then in the circumstances just described, the frequencies of $AA$, $Aa$ and $aa$ are $p^2$, $2pq$, and $q^2$, respectively.

Consider another physical realization of the simple mathematical idea involved in the Hardy–Weinberg Law. A shoe manufacturer produces brown shoes and black shoes. By accident, the assembly line has not kept the shoes together in pairs, but has dumped all the left shoes into one pile and all the right shoes into another. The shoe manufacturer wants to know what the result will be if a machine randomly samples from the two piles and assembles pairs of shoes. If $p$ is the frequency of black shoes and $q$ is the frequency of brown ones in each pile, then the expected frequency of the three possible pairs will be $p^2$, $2pq$, and $q^2$. Many other examples of this mathematical sort could be described.

Suppose we applied the Hardy–Weinberg Law to a population of *Drosophila*. These fruit flies are biological objects; they are alive and the Hardy–Weinberg Law describes an important fact about how they reproduce. As just noted, the same mathematical structure can be applied to shoes. But shoes are not alive; the process by which the machine forms pairs by random sampling is not a biological one.

I wish to introduce a piece of terminology: the Shoe/Fly Fallacy is the mistaken piece of reasoning embodied in the following argument:

Flies are alive.
Flies are described by law L.
Shoes are described by law L.
Hence, shoes are alive.

A variant of this argument focuses on a specific biological property—like reproduction—rather than on the generic property of "being alive."

Functionalist theories abstract away from physical details. They go too far—confusing mathematical form with biological (or psychological) subject matter—when they commit the Shoe/Fly Fallacy. The result is an overly liberal conception of life (or mind).

The idea of the Shoe/Fly Fallacy is a useful corrective against overhasty claims that a particular artificial system is alive or exhibits some range of biological characteristics. If one is tempted to make such claims, one should try to describe a system that has the relevant formal characteristics but is clearly not alive. The Popperian attitude of attempting to falsify is a useful one.

Consider, for example, the recent phenomenon of computer viruses. Are these alive? These cybernetic entities can make their way into a host computer and take over some of the computer's memory. They also can "reproduce" themselves and undergo "mutation." Are these mere metaphors, or should we conclude that computer viruses are alive?

To use the idea of the Shoe/Fly Fallacy to help answer this question, let's consider another, similar system that is not alive. Consider a successful chain letter. Because of its characteristics, it is attractive to its "hosts" (i.e., to the individuals who receive them). These hosts then make copies of the letters and send them to others. Copying errors occur, so the letters mutate.

I don't see any reason to say that the letters are alive. Rather, they are related to host individuals in such a way that more and more copies of the letters are produced. If computer viruses do no more than chain letters do, then computer viruses are not alive either.

Note the "if" in the last sentence. I do not claim that computer viruses, or something like them, cannot be alive. Rather, I say that the idea of the Shoe/Fly Fallacy provides a convenient format for approaching such questions in a suitably skeptical and Popperian manner.

## FROM HUMAN COGNITION TO AL, FROM BIOLOGY TO AL

One of the very attractive features of AI is that it capitalizes on an independently plausible thesis about human psychology. A fruitful research program about human cognition is based on the idea that cognition involves computational manipulations of representations. Perhaps some representations obtain their intentionality via a connection between mind and world. Once in place, these representations give rise to other representations by way of processes that exploit the formal (internal) properties of the representations. To the degree that computers can be built that form and manipulate representations, to that degree will they possess cognitive states.

Of course, very simple mechanical devices form and manipulate representations. Gas gauges in cars and thermostats are examples, but it seems entirely wrong to say that they think. Perhaps the computational view can explain why this is true by further specifying which kinds of representations and which kinds of computational manipulations are needed for cognition.

Can the analogous case be made for the artificial life research program? Can an independent case be made for the idea that biological processes in naturally occurring organisms involve the formation and manipulation of representations? If this point can be defended for naturally occurring organisms, then to the degree that computers can form and manipulate representations in the right ways, to that degree will the AL research program appear plausible.

Of course, human and animal psychology is part of human and animal biology. So it is trivially true that some biological properties involve the formation and manipulation of representations. But this allows AL to be no more than AL. The question, therefore, should be whether life processes other than psychological ones involve computations on representations.

For some biological processes, the idea that they essentially involve the formation and manipulation of representations appears plausible. Thanks to our understanding of DNA, we can see ontogenesis and reproduction as processes that involve representations. It is natural to view an organism's genome as a set of instructions for constructing the organism's phenotype. This idea becomes most plausible when the phenotypic traits of interest are relatively invariant over changes in the environment. The idea that genome represents phenotype must not run afoul of the fact that phenotypes are the result of a gene/environment interaction. By the same token, blueprints of buildings don't determine the character of

a building in every detail. The building materials available in the environment and the skills of workers also play a role. But this does not stop us from thinking of the blueprint as a set of instructions.

In conceding that computers could exemplify biological characteristics like development and reproduction, I am not saying that any computers now do. In particular, computers that merely manipulate theories about growth and reproduction are not themselves participants in those processes. Again, the point is that a description of a bridge is not a bridge.

What of other biological properties and processes? Are they essentially computational? What of digestion? Does it involve the formation and manipulation of representations? Arguably not. Digestion operates on food particles; its function is to extract energy from the environment that the system can use. The digestive process, *per se,* is not computational.

This is not to deny that in some systems digestion may be influenced by computational processes. In human beings, if you are in a bad mood, this can cause indigestion. If the mental state is understood computationally, then digestion in this instance is influenced by computational processes. But this effect comes from without. This example does not undermine the claim that digestion is not itself a computational process.

Although I admit that this claim about digestion may be wrong, it is important that one not refute it by trivializing the concepts of representation and computation. Digestion works by breaking down food particles into various constituents. The process can be described in terms of a set of procedures that the digestive system follows. Isn't it a short step, then, to describing digestion as the execution of a program? Since the program is a representation, won't one thereby have provided a computational theory of digestion?

To see what is wrong with this argument, we need to use a distinction that Kant once drew in a quite different connection. It is the distinction between following a rule and acting in accordance with a rule. When a system follows a rule, it consults a representation; the character of this representation guides the system's behavior. On the other hand, when a system acts in accordance with a rule, it does not consult a representation; rather, it merely behaves as if it had consulted the rule.

The planets move in ellipses around the sun. Are they following a rule or are they just acting in accordance with a rule? Surely only the latter. No representation guides their behavior. This is why it isn't possible to provide a computational theory of planetary motion without trivializing the ideas of computation and representation. I conjecture that the same may be true of many biological processes; perhaps digestion is a plausible example.

In saying that digestion is not a computational process, I am not saying that computers can't digest. Planetary motion is not a computational process, but this does not mean that computers can't move in elliptical orbits. Computers can do lots of things that have nothing to do with the fact that they are computers. They can be doorstops. Maybe some of them can digest food. But this has nothing to do with whether a computational model of digestion will be correct.

I have focused on various biological processes and asked whether computers can instantiate or participate in them. But what about the umbrella question? Can computers be alive? If Turing's question was whether computers can think, shouldn't the parallel question focus on what it is to be alive, rather than on more fine-grained concepts like reproduction, growth, selection, and digestion?

I have left this question for last because it is the fuzziest. The problem is that biology seems to have little to tell us about what it is to be alive. This is not to deny that lots of detailed knowledge is available concerning various living systems. But it is hard to see which biological theories really tell us about the nature of life. Don't be misled by the fact that biology has lots to say about the characteristics of terrestrial life. The point is that there is little in the way of a principled answer to the question of which features of terrestrial life are required for being alive and which are accidental.

Actually, the situation is not so different in cognitive science. Psychologists and others have lots to tell us about this or that psychological process. They can provide information about the psychologies of particular systems. But psychologists do not seem to take up the question "What is the nature of mind?" where this question is understood in a suitable, nonchauvinistic way.

Perhaps you are thinking that these are the very questions a philosopher should be able to answer. If the sciences ignore questions of such generality, then it

is up to philosophy to answer them. I am skeptical about this. Although philosophers may help clarify the implications of various scientific theories, I really doubt that a purely philosophical answer to these questions is possible. So if the sciences in question do not address them, we are pretty much out of luck.

On the other hand, I can't see that it matters much. If a machine can be built that exemplifies various biological processes and properties, why should it still be interesting to say whether it is alive? This question should not preoccupy AL any more than the parallel question should be a hang up for AI. If a machine can perceive, remember, desire, and believe, what remains of the question of whether it has a mind? If a machine can extract energy from its environment, grow, repair damage to its body, and reproduce, what remains of the issue of whether it is "really" alive?

Again, it is important to not lose sight of the if's in the previous two sentences. I am not saying that it is unimportant to ask what the nature of mind or the nature of life is. Rather, I am suggesting that these general questions be approached by focusing on more specific psychological and biological properties. I believe that this strategy makes the general questions more tractable; in addition, I cannot see that the general questions retain much interest after the more specific ones are answered.

## CONCLUDING REMARKS

Functionalism, both in the study of mind and the study of life, is a liberating doctrine. It leads us to view human cognition and terrestrial organisms as examples of mind and life. To understand mind and life, we must abstract away from physical details. The problem is to do this without going too far.

One advantage that AL has over AI is that terrestrial life is in many ways far better understood than the human mind. The AL theorist can often exploit rather detailed knowledge of the way life processes are implemented in naturally occurring organisms; even though the goal is to generalize away from these examples, real knowledge of the base cases can provide a great theoretical advantage. Theorists in AI are usually not so lucky. Human cognition is not at all well understood, so the goal of providing a (more) general theory of intelligence cannot exploit a detailed knowledge of the base cases.

An immediate corollary of the functionalist thesis of multiple realizability is that biological and psychological problems are not to be solved by considering physical theories. Existing quantum mechanics is not the answer, nor do biological and psychological phenomena show that some present physical theory is inadequate. Functionalism decouples physics and the special sciences. This does not mean that functionalism is the correct view to take for each and every biological problem; perhaps some biological problems are physical problems in disguise. The point is that, if one is a functionalist about some biological process, one should not look to physics for much theoretical help.

The Turing Test embodies a behavioristic criterion of adequacy. It is not plausible for psychological characteristics, though it may be correct for a number of biological ones. However, for virtually all biological processes, the behaviors required must be rather different from outputs printed on a tape. A desktop computer that is running a question/answer program is not reproducing, developing, evolving, or digesting. It does none of these things, even when the program describes the processes of reproduction, development, digestion, or evolution.

It is sometimes suggested that, when a computer simulation is detailed enough, it then becomes plausible to say that the computer is an instance of the objects and processes that it simulates. A computer simulation of a bridge can be treated as a bridge, when there are simulated people on it and a simulated river flowing underneath. By now I hope it is obvious why I regard this suggestion as mistaken. The problem with computer simulations is not that they are simplified representations, but that they are representations. Even a complete description of a bridge—one faithful in every detail—would still be a very different object from a real bridge.

Perhaps any subject matter can be provided with a computer model. This merely means that a description of the dynamics can be encoded in some computer language. It does not follow from this that all processes are computational. Reproduction is a computational process because it involves the transformation of representations. Digestion does not seem to have this characteristic. The AL research program has plenty going for it; there is no need for overstatement.

## NOTES

This chapter originally appeared in Christopher G. Langton, Charles Taylor, J. Doyne Farmer, and Steen Rasmussen (Eds.), *Artificial life II,* pp. 749–765, Boulder, CO; Oxford: Westview Press, 2003.

## REFERENCES

1. Block, N. (1975). Troubles with functionalism. In W. Savage (Ed.), *Perception and cognition: Issues in the foundations of psychology* (Minnesota studies in the philosophy of science, volume IX) (pp. 261–326). Minneapolis: University of Minnesota Press.
2. Block, N. (1981). Psychologism and behaviorism. *Philosophical Review*, **90**, 5–43.
3. Dennett, D. (1978). *Brainstorms.* Cambridge, MA: MIT Press.
4. Dretske, F. (1985). Machines and the mental. *Proceedings and Addresses of the American Philosophical Associations*, **59**, 23–33.
5. Fisher, R. (1958). *The genetical theory of natural selection*. New York: Dover.
6. Fodor, J. (1968). *Psychological explanation*. New York: Random House.
7. Lewontin, R. (1970). The units of selection. *Annual Review of Ecology and Systematics*, **1**, 1–14.
8. Pattee, H. (1989). Simulations, realizations, and theories of life. In C. G. Langton (Ed.), *Artificial life* (Santa Fe Institute studies in the sciences of complexity, proceedings volume VI) (pp. 63–77). Redwood City, CA: Addison-Wesley.
9. Penrose, R. (1990). *The emperor's new mind*. Oxford: Oxford University Press.
10. Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology*, **47**, 44–50.
11. Putnam, H. (1975). The nature of mental states. In H. Putnam, *Mind, language, and reality* (pp. 429–440). Cambridge, UK: Cambridge University Press.
12. Ryle, G. (1949). *The concept of mind*. New York: Barnes & Noble.
13. Searle, J. (1980). Minds, brains, and programs. *Behavior and Brain Sciences*, **3**, 417–457.
14. Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review*, **68**, 141–156.
15. Sober, E. (1985). Methodological behaviorism, evolution, and game theory. In J. Fetzer (Ed.), *Sociobiology and epistemology* (pp. 181–200). Dordrecht: Reidel.
16. Sober, E. (1984). *The nature of selection*. Cambridge, MA: MIT Press.
17. Turing, A. (1950). Computing machinery and intelligence. *Mind*, **59**, 433–460.
18. Wittgenstein, L. (1953). *Philosophical investigations.* Oxford: Basil Blackwell.