Brian L. Keeley
Experimental Philosophy Laboratory, Department of Philosophy (0302), University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0302; e-mail: bkeeley@ucsd.edu

Against the Global Replacement: On the Application of the Philosophy of Artificial Intelligence to Artificial Life

This paper considers itself a complement to the recent wealth of literature suggesting a strong philosophical relationship between artificial life (AL) and artificial intelligence (AI). I seek to point out areas where this analogy seems to break down or where it would lead us to draw hasty conclusions about the philosophical situation of AL. First, I sketch a thought experiment (based on the work of Tom Ray) that purports to suggest how such experiments should be evaluated. In doing so, I suggest that treating AL experiments as if they were just AI experiments applied to a new domain may lead us to see problems (like Searle's "Chinese room") that just aren't there. In the second half of the paper, I take a look at the reasons behind suggesting there is a philosophical relationship between the two fields. I characterize the strong thesis for a translation of AI concepts, metaphors, and arguments into AL as the "global replacement strategy." Such a strategy is only fruitful in as much as there is a strong analogy between AI and AL. I conclude the paper with a discussion of two areas where such a strong analogy seems to break down. These areas relate to eliminative materialism and the lack of a "subjective" element in biology. I conclude

Artificial Life III, Ed. Christopher G. Langton, SFI Studies in the Sciences of Complexity, Proc. Vol. XVII, Addison-Wesley, 1994

that the burden of proof lies with one who wishes to import a concept from another discipline into AL, even if that other discipline is AI.

1. INTRODUCTION

In many ways, artificial life (AL) has long been the poor, younger sibling of artificial intelligence (AI). The two fields share many superficial similarities: Where AI can be seen as the synthetic, engineering side of the more analytic theoretical psychology, AL can be seen as the synthetic, engineering side of the more analytic theoretical biology. Both fields make extensive use of the modern digital computer, currently only as *models* but also, practitioners in both fields hope, potentially as *instances* or *examples* of the phenomena they study. The philosophical literature of AL is littered with concepts, metaphors, and arguments taken from AI. Variously, there is mention of AL Turing tests, AL dualism, AL functionalism, AL Chinese rooms, etc., all of which are concepts familiar from decades of discussion in AI.

Some, like Eliot Sober,[21] have even gone as far as to point to a strong analogy between AI and AL, an analogy that seems to vindicate such wholesale philosophical looting of traditional positions in AI. But it is the nature of analogies—even strong analogies—that there are differences between the two related entities. AL *is not* AI. On the basis of these differences, I argue that artificial life would be best served by originating new philosophical positions and metaphors of its own, without haphazardly borrowing such constructions from artificial intelligence. The spirit of this paper is to act as a complement to the growing pool of literature which either documents or implies similarities between AL and AI. Instead, I will try to highlight the dissimilarities between the two endeavors. In particular, I wish to point out areas where these differences are actually advantageous for AL, and where looking at AL through "AI-colored glasses" will lead one to see problems that may not be there.

The paper begins with a thought experiment, which is meant to capture an idealized picture of one of the goals of AL: to create life in a computer. Based loosely on the work of Tom Ray,[18] it is intended to explore the relationship between natural systems and AL programs that purport to exhibit biological phenomena. I hope to determine *the basis* on which we should make the decision of whether a given AL experiment is a genuine example of genuine *artificial* life. In doing so, I suggest the bases for this judgement are different from those traditionally involved in determining whether a system is an example of artificial intelligence. I conclude that treating AL as if it were just AI applied to different natural phenomena leads one to grapple with "Chinese room" objections to AL. However, I argue that the evaluation of AL experiments are sufficiently different to allow them to escape such considerations.

I then turn to the more abstract issue of the proposed analogy between AI and AL. What are the arguments in its favor? More importantly, given such an analogy

what license does it give when deciding which concepts and metaphors from AI should be taken up in AL? I argue on the side of caution when "translating" the philosophy of AI into a philosophy of AL, pointing out that doing this properly requires a familiarity with *both* what is analogous and disanalogous between the fields. With this in mind, I end the paper with a discussion of two strong disanalogies between AI and AL: the lack of a viable eliminative materialist position within AL and the lack of anything analogous to the "problem of consciousness" in AL.

2.1 BLOB WORLD VS. BLIP WORLD: AN ARTIFICIAL LIFE METAPHOR

Let us now turn to that old chestnut of philosophical methodology, the thought experiment. In the following, I will consider an idealized example of an AL experiment in order to examine where the epistemological priorities lie, and whether they lie in places where a strong relationship to AI would suggest.

Imagine, if you will, a medium that exhibits some phenomena of interest to biology (Figure 1(a)). Unfortunately, the scale of these phenomena is quite microscopic—it is invisible to the naked eye—requiring the use some kind of "visualizer" that can magnify and regularize the behavior so that it can be seen on a CRT screen. On that screen we see an image consisting of slowly moving circles and some darker masses, all embedded within a heterogeneous medium. As we watch, some of the circles envelope the dark masses, while other circles occasionally split into two more-or-less identical circles. Let us call this medium and its phenomena "blob world."

But now imagine another medium, which also exhibits some interesting behavior (Figure 1(b)). It too is very small and otherwise invisible to the naked eye, so another kind of "visualizer" is required to make the phenomena visible on a CRT screen. What we see on this screen is a column of letters: "0080aaa," "0045aab," "0061acc," etc., next to that are some horizontal bars that are hectically pulsing out and back across the screen. As we watch, new letter combinations come into existence, while others disappear. Though the appropriateness of doing so is not yet apparent, let us call this second medium and its phenomena "blip world."

It should be no surprise when a microbiologist comes around and tells us that blob world is a group of microscopic single-celled organisms feeding and multiplying in a petri dish. And, as she has been recently reading up on the doings in AL, she also tells us that the blip-world output looks a lot like the real-time output of Tom Ray's *Tierra* simulator.[18] (Blip world is not identical to Tierra in all its details—Blip world is simplified for ease of presentation—but they are meant to be identical in their philosophical status. In this sense, Tierra is one of a variety of possible blip worlds. Given the success Ray has had with Tierra, it seems reasonable to expect to see similar research programs in the future.)

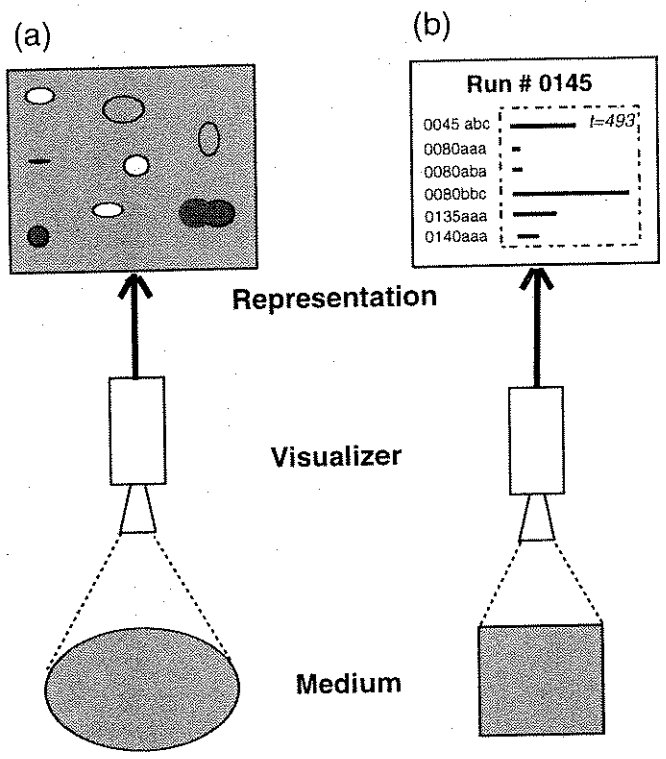

FIGURE 1 (a) First look at Blob world. (b) Blip world.

In blip world, each alpha-numeric string identifies an artificial "organism" which, in turn, is a bit of machine-level computer code. The bar next to the identifier represents the proportion of memory occupied by the token instances of that type of code. Each "organism" is essentially a piece of self-replicating code which contains the instructions required for replicating itself in the medium of RAM. As such, the code is both genotype and phenotype; it is both the instructions for replicating and what is replicated. If allowed, just one of these bits of code would soon replicate itself to the point that it filled up the entire memory with little copies of itself. However, this is prevented by two mechanisms. First, the code is not allowed to replicate itself perfectly. Every now and then it messes up and writes a "0" instead of a "1" or vice versa. In this way, *mutations* of the initial seed code enter the population. Just as with natural organisms, most of these mutations are fatal, in that they do not lead to code capable of self-replication, but some do turn out to be viable in this sense. Second, in order to keep the successfully replicating code from overrunning the system, a proportion (determined by age) is culled each generation. We can imagine that blip world exhibits the same interesting behavior as

Tierra, including "parasites" that locate themselves next to "hosts" and trick these hosts into copying the parasite's code instead of their own, and "hyperparasites" that play a similar trick on the parasites. We also see extended periods of stasis in the pool of different types of code interspersed with spurts of tumultuous change as new types compete with and replace the old.

Blip worlds like Tierra are AL simulations. We are called upon to evaluate the claim that what is going on in these worlds is similar enough to what is going on in *real* biological systems, such as the petri dish, that the predicate "alive" or "biological" ought to be applied to each with equal force. In essence, the claim is that *blip world contains life*, just as biologists agree that blob world does. The only relevant difference, so goes the claim, is that blip world exhibits manmade or *artificial* life, whereas *natural* life is going on in blob world. The only relevant difference between the two situations is one of origins. To evaluate this claim, the two kinds of systems need to be scrutinized in order to determine any relevant dissimilarities or asymmetries between the two situations. It should be kept in mind that the task here is not to determine *whether the claim is true*, but to say *in virtue of what* it is or is not. This latter task is the philosophical one that we must confront. Only after we have determined the basis on which the decision of "life" or "not life" is to be made, can we turn to specific details of a specific system (like Tierra) and attempt to make the decision.

The first difference in these two scenarios is in what is displayed on the screens. With blob world, we see a picture of the petri dish, whereas with blip world we see some kind of data chart. It is like the difference between seeing William S. Burroughs through the lense of a video camera and reading his biography. Clearly, one feels, the two scenarios must be markedly different. In blob world, "real" eating and reproducing is going on. We actually can see it on the screen. But in blip world, we see some kind of symbol manipulation and are treated to the results of these computations on its output screen. At best, only simulated—*as if*—eating and reproducing is going on.

However, this conclusion is hasty. The behavior of the two scenarios are indeed visualized differently. But this is due primarily to the different temporal scales of the two situations. Let us call the representation given in blob world a *window* representation (WR), an as-accurate-as-possible representation of the "look" of the system. It provides the viewer with a "window" on the medium. It is what we imagine we would see if we were miniaturized, or if the petri dish and all of its inhabitants were magically enlarged to the size of a swimming pool. Let us call the representation of blip world a *dynamic time-course* representation (DTCR): a representation of the long-term, gross dynamics of the system represented in aggregate, statistical form.

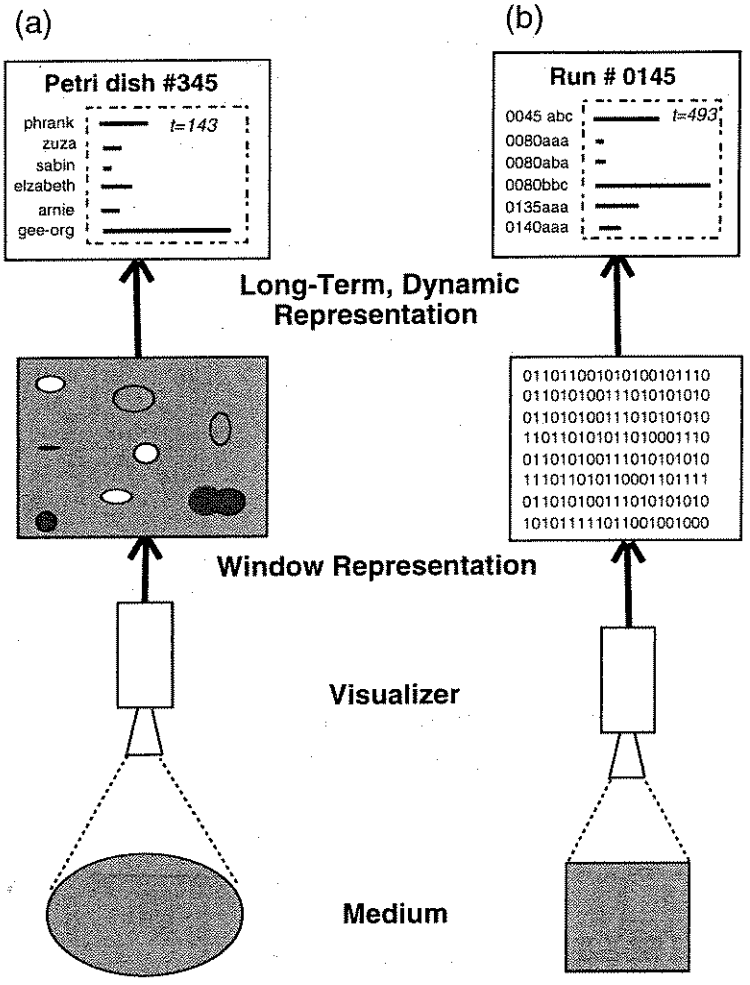

FIGURE 2 (a) A more complete look at blob world and (b) blip world.

If these representations were somehow uniquely and exclusively tied to the systems at hand, this indeed would be an important difference between them. But that one of these representations is commonly and preferentially used with each respective system is just an artifact of what we find most informative about each system and which representation is easiest to generate. For instance, a DTCR of blob world could be generated by identifying the individual cells and keeping track of their movement and reproduction (Figure 2(a)). In other words, if we were patient enough, we could keep track of lineages (perhaps by chemical tagging) and the

percentage of the blob world that each lineage occupied. Admittedly it would be difficult to generate such data (especially in real time), but there is no reason *in principle* why it could not be done.

A critical point is that the dynamics displayed in the DTCRs of blip world and blob world show similarities. On the basis of this kind of similarity alone, we are led to believe that AL is possible. Artificial life's biggest claim to fame is that computer models of biological systems are often remarkably good at capturing the gross, high-level dynamics of biological systems. The literature is packed with computer models that capture population dynamics, the evolution of cooperative behavior, speciation, learning, etc. Often, what these models capture are such examples of the "look and feel" of biological systems, but some systems, in particular blip worlds, capture more.

Just as a DTCR of blob world can be produced, it is fairly trivial to produce a WR of blip world (Figure 2(b)). It would be a bit map of memory: a plane of 1's and 0's blinking on and off at a very high rate. These numbers represent the patterns of high and low voltages present in the memory of the computer. Where the WR of the petri dish is made up of "blobs," the WR of this alleged "electronic petri dish" would be made up of "blips." A blip world WR would be pretty meaningless to most viewers, and this is why this representation is rarely used to display the behavior of blip-world systems like Tierra.

But the patterns *are there to be seen*, if one could train oneself to see them. Properly trained, one would see certain strings of bits are more numerous than others, as the more successful codes (and their "children") copied themselves. If one watched closely enough, new types would be seen arising in the population, as mutation and selection occurs. The existence of these patterns is another crucial similarity between blip world and blob world and, as discussed below, this similarity is lacking or unimportant for similarly constructed AI models.

The asymmetry in the representational forms is then an accident of the combined effects of the dynamics of the systems involved, the limits of our perceptual capabilities, and our familiarity with types of representations. It is more informative to see the time-course data of blip world, and it is relatively easy to generate. With life in Petri dishes, such aggregate data is hard to produce. Also familiarity with blob world WRs makes it easier to see the behavior in which we are interested using that kind of representation.

2.2 WHAT OUGHT TO BE MADE FROM THIS METAPHOR?

In many ways then, the situations with blip world and blob world are analogous. There are indeed differences between them: (a) blip world seems to behave at a much higher speed and (b) behavior in blip world is much more easily quantified than that of blob world. We might also add (c) the form of energy used by both systems is different as well—blip-world organisms use electricity where blob-world organisms use sugars and sunlight. However, one feels that these differences are not

relevant to the question of whether blip world is truly biological. We can imagine genuine living systems whose metabolisms and life cycles occurred at a much higher rate than that of life on earth. We can imagine having developed the technical know-how to produce measuring devices capable of producing from petri dishes the kinds of DTCRs we can generate so easily for blip world. Similarly, the details of how living systems convert energy into useful behavior also seems to be accidental and not an essential property of life. That blip world is different on these counts only illustrate that, if it is truly biological, it will be a *different* biology from life-as-we-know-it.

Some might say that so far I have overlooked an important difference between the two systems. It might be argued that blip world is "merely a simulation," that all that is going on in blip-world systems is mere symbol manipulation. The crux of this complaint can be traced to John Searle's now classic 1980 paper, "Minds, Brains, and Programs,"[19] in which he claims to refute what he calls "strong artificial intelligence." Strong artificial intelligence (AI) is the claim that an appropriately programmed computer can be an instance of a truly conscious, intelligent system.

Searle carries out this refutation through the use of his "Chinese room" thought experiment, which purports to show that mere rule following and symbol manipulation is not sufficient for true understanding, meaning, or intentionality. A conclusion Searle draws from his arguments is that the best that an AI program could ever be is a *model* or *simulation* of meaningful behavior, but never an *instance* of it. For the purposes of this paper, I am going to accept what I feel is a major conclusion of Searle's argument: a system cannot be said to exhibit a property such as "intelligence" (or, in the case of AL, "life") by virtue of its computational properties alone. As Searle might put it, computational properties are not the proper kind of causal properties to instantiate real intelligence (or life).

In his paper in these proceedings, Stevan Harnad has suggested just such an application of Searle's argument to the endeavor of artificial life.[13] Specifically, his claim is that, unless it is *grounded* (hooked up to the world with sensors and effectors), the best that an AL computer program could ever be is a *simulation* of life, never an *instance* of it. It would seem that such a criticism is supposed to apply to blip-world programs like Tierra. (However, it is difficult to be sure, as he never mentions any specific AL research by name.) Blip world is not hooked up to the world outside the computer in any way significantly different from the way in which traditional AI models are. Apparently, we are invited to draw similar conclusions about the reality of such AL models as Searle and Harnad draw about such AI models.

However, to draw this quick conclusion is to fall into the trap of looking at AL models as if they are simply AI models applied to a different domain. The two situations certainly look alike: a computer program crunching away on a program and throwing data up on a screen that bears a striking resemblance to what some natural phenomena would throw up on a screen and, if that resemblance is close enough, concluding that the computer is instantiating that natural phenomena as

well. And, if Searle has refuted this argument in AI, surely he has done so in AL, as well.

Wrong. To see why, consider how the position known as "functionalism" is used in AI. Particularly in its original Turing test form, functionalism embodies the claim that, to some degree of abstraction, what a system is made of does not matter in the determination of whether it is "intelligent," "conscious," "intentional," etc. All that matters is whether it *behaves* in the correct way. The Turing test[23] sets out a strict procedure for determining what is legitimate behavioral evidence for making this judgement: answers to questions input to the system via a teletype. Modern versions of functionalism substitute other behaviors in place of those of the Turing test, such as Harnad's suggestion that the system be able to sense the world and execute robotic behaviors, but the multiple realizability thesis is maintained by throwing out of court any evidence based directly upon *how* the system produces its behavior. (The "multiple realizability thesis" is the claim that, in some sense, what a system is made out of is irrelevant to whether it is an instance of some phenomenon.) In traditional AI functionalism, we are called upon to determine whether a system's gross behavioral output meets some criteria, and then an attribution ("intelligent," "conscious") is *projected back* onto the specific physical system that generated the behavior.

However, this is not how the claim of "life" is decided in the case of blip worlds. I argue that whether blip world contains living things is *not* determined on the basis of what is displayed in either WR or DTCR, as would be the case if blip world were an AI system. *Blip world is evaluated as living or not on the basis of what behavior it exhibits in the medium, the RAM.* If the behavior of the medium is sufficiently like that of the Petri dish, then we call it biological, living, or whatever. Given that both of the media are not directly observable, and given the confidence that the production of the representations does not introduce any artifacts, then the evaluation, in practice, will be carried out by comparing the behavior of these representations. But it should always be clear that the both the WR and the DTCR only can be considered "lifelike" in virtue of the lifelike behavior of the medium that gives rise to them. In AL, the physical medium is judged to be lifelike or not and then that attribution is projected *forward* (not *backward*, as in AI) to the representations that system generates.

If there is an analogous position in the philosophy of AI for this position, it would be what Andy Clark[4] has dubbed *microfunctionalism*. This position, developed as a defense of the more biologically plausible approach to AI known as *connectionism*, reels in the more free-ranging liberalism of traditional functionalism, arguing that it *does* matter what a system is made of and how it produces its surprising behavior. According to microfunctionalism, evidence about the mechanisms utilized by a system to generate its behavior can be legitimately used to determine whether a system should be attributed with a property, such as "intelligence" or "consciousness."

In a similar way, microfunctionalism for AL argues that it is crucial to look inside the computer—at the behavior of the medium—to see what behavior there

is and how it is produced. If, upon looking into the system, variables and sets of rules (in other words, symbols and procedures for manipulating those symbols) are found, then it can be concluded (with Searle and Harnad) that the system is merely a simulation of life. If instead a system that exhibits physical properties relevantly similar to those of "real" living systems is discovered, then we may conclude that there is indeed life in blip world.

Indeed it may turn out to be the case that we will decide that what is going on in the RAM of a specific blip-world system like Tierra is just not similar enough to natural life to warrant the claim of artificial *life*. For instance, the Tierra organisms lack both development (they lack anything that resembles morphogenesis) and metabolism, and biologists may decide that these features are indeed crucial to its characterization as a true biological system. And perhaps, as Michael Dyer (in conversation) has suggested, the physics of the internal world of a computer is just too simple and regular, compared to that of the Terrestrial world, in which life as we know it has developed, to support the complex entities typically associated with life. However, such a decision must be made primarily on the basis of continuing work within theoretical biology.

In any case, I have proposed an answer to the philosophical question I set out when I introduced the blip-world vs. blob-world thought experiment: When deciding whether a particular blip-world program is truly biological, in virtue of what is that decision made? I have argued that it is in virtue of blip world's *physical* properties (not its *computational* properties) that it exhibits relevantly biological behavior. While it is true that the medium in which this behavior is found is a "computer," we should never forget that our computer is not some kind of Platonic "purely computational system"; it is a very down-to-earth physical system, a machine. Not everything that a computer can do is "computational" in nature. My NeXT computer workstation can not only simulate a paperweight, it can actually instantiate one as well. It can not only simulate the heat output of a NeXT workstation, as a NeXT workstation it also produces real (not simulated) heat.

The claim here is that what is going on inside a computer running a blip-world program is *not* a computational simulation of life. It is an automated physical procedure for seeding the computer's memory with appropriate physical patterns of high and low voltages, and for appropriately visualizing the resulting dynamics. The fact that all this is going on in a medium that is typically used to perform operations that are interpretable systematically in computational terms is irrelevant.

This claim is highly counterintuitive. I am suggesting that if blip world is judged to be alive, it will be on the basis of its physical, not its computational, properties. The blip world I have described exhibits the property of self-replication in the same way that my workstation exhibits the property of producing heat. Real, physical self-replication is going on inside the computer's RAM, as certain patterns of high and low voltages manipulate neighboring locations until they exhibit an identical pattern of high and low voltages. This is not simulated or *as if* self-replication; this is *instantiated* self-replication!

(Note that I are not claiming that the simulation is not occurring on a computer [this is obvious], I am only claiming that such a simulation is not making use of the well-known computational properties of the computer. Similarly, the use of my NeXT computer to determine the heat output of an identical make of computer is to use a computer as non-computational simulation. Perhaps the term "model" [as in the scale models used by architects, or the models built by children] is a better term for this situation.)

I hope that I have shown that the application of traditional AI philosophical analysis to *prima facie* similar situations in AL can be misleading, saddling AL with problems and concerns (like the Chinese room) that it can do well without. However, the application of AI thinking to AL is an appealing one, and presumably has its utility, as in the case of microfunctionalism. To what extent, and in which situations, is such a comparison fruitful? This question is the topic of the second half of this paper.

3. ANALOGIES AND STRATEGIES

In "Learning from Functionalism—The Prospects for Strong Artificial Life," Elliot Sober[21] explores the following analogy: "Artificial intelligence is to psychology as artificial life is to biology." With this analogy (which I call the *Sober analogy*) he sketches a variety of positions and concerns from the traditional philosophy of AI as they would appear in the philosophy of AL. He discusses "strong" and "weak" AL, biological dualism and identity theory, biological multiple realizability, etc. Sober eventually argues for a functionalist approach to biology and AL that parallels the prominent philosophical position of the same name found in psychology and AI.[1]

Sober is not alone in seeing parallels between AI and AL. In his seminal essay introducing the first AL proceedings, Chris Langton[16] follows a similar path (see, in particular, Figure 11, p. 40). There he notes a similarity between connectionist AI (where relatively complicated "intelligent" behavior is generated using a relatively simple structural substrate as in connectionism) and AL modeling (where relatively complicated "living" behavior is generated using a relatively simple substrate as in cellular automata). Similarly, in that same volume, Pattee[17] notes, "It is clear from this workshop [Artificial Life I] that artificial life studies have closer roots in artificial intelligence and computational modeling than in biology itself."

While this evidence indicates a connection between AL and AI, Sober is arguing for a close relationship between the philosophical situations of each field.

[1] Traditionally, the philosophy of AI has been seen as a specialization of the more general set of concerns of the philosophy of psychology. Similarly, one would expect that a "philosophy of AL" would be a specialization of the philosophy of biology. In lieu of the cumbersome phrasing "philosophies of psychology and AI" and "philosophies of biology and AL," I will refer to only the "philosophy of AI" and the "philosophy of AL" for the sake of brevity.

While Sober argues for an AL version of functionalism, others discuss an AL "Turing test,"[1] an AL "Chinese room,"[13,15] and an AL hardware-software distinction.[8] Given that AL is generally free of philosophical discussion (some would say *refreshingly* free), these examples suggest that Sober is not alone in pointing out a deep philosophical connection between AI and AL.

The Sober analogy is an appealing one, and there is no doubt a lot of truth in it. Where AI is the synthetic, engineering counterpart of the more analytic science of theoretical psychology, AL is the synthetic, engineering counterpart of the more analytic science of theoretical biology. Both AI and AL make extensive use of the digital computer and computer models of their respective phenomena. Both AI and AL argue that there is no reason why we can't build artificial examples of what have been phenomena of purely natural origin.

However, analogies are not incredibly useful by themselves. They just suggest that there are similarities (and differences) between two things. A methodology based on analogy would be more useful. One wants to turn a simple logical relationship into a methodology. For a new endeavor like AL, such a methodology might include setting out the set of important philosophical metaphors, positions, and distinctions to be used in that endeavor. I feel that in the above examples—the application of traditional AI distinctions to AL—imply just such a methodology. In its most extreme form, this implicit strategy (which I call the *Global Replacement Strategy*, or GRS for short) involves taking thirty years of avid discussion in the philosophy of AI and translating it into what then will be the "philosophy of AL." This strategy, apparently supported by the Sober analogy, gives AL a way of generating a complete and well-worked-out philosophical landscape, merely by taking the canon of the philosophy of AI and (stealing a concept from word processing) *globally replacing* all occurrences of "intelligence" with "life."

This extreme application of the Sober analogy is not without its merits. It allows the still-embryonic AL to take advantage of the large philosophical armory that AI has struggled to develop over the better part of three decades. AL can dispense with doing any of this hard work for itself. In a mere five years since its inception, so goes the GRS argument, Sober has given AL a rich and varied philosophical tapestry of positions, arguments, and metaphors to rival that of any other, more established endeavor.

However, no matter how appealing it might seem, GRS is not the best course for the AL community to take. There is good reason to believe that there is much to be gained by originating a novel philosophy of AL, with little derivation from traditional philosophies of psychology and AI. As illustrated in the blip-world vs. blob-world example, thinking of AL in traditional AI terms can lead one one astray. This example illustrates the dangers of the GRS, but a more general account of its hazards is needed.

4.1 DISANALOGIES BETWEEN LIFE AND MIND

Like all such analogies, the Sober analogy does not claim that the central phenomena of psychology ("mind") and biology ("life") are identical, but it does suggest that the way that these phenomena are (or should be) handled in their respective domains are significantly parallel. GRS is calculated to use that parallelism, turning it into a constructive strategy for defining the proper philosophical problem space of AL. But while attractive, there is a problem with this picture. The existence of any important disanalogies between the domains of psychology and biology would point to large areas of concerns that would resist the simple translation of one field into the other. In the remainder of this paper, I will consider what I believe are the two most important differences between the phenomena of life and mind: the lack of a strong eliminative materialist position in biology and the lack of a biological concern with the subjective.

4.2 FOLK BIOLOGY AND ELIMINATIVE MATERIALISM

When looking at the arguments of those who wish to allege a strong analogy, it is often more instructive to note what the author *fails* to mention, rather than what he actually does. Among the positions traditionally available to the philosopher of AI (and, *mutatis mutandi*, to the would-be philosopher of AL), Sober mentions dualism, identity theory, and functionalism, among others. But one position he fails to mention is *eliminative materialism* (EM). Originally argued by Paul Feyerabend[9,10,11] and currently championed by Stephen Stich[22] and Paul M. Churchland,[2,3] EM is primarily a thesis about proper scientific explanation. In particular, it seeks to reject the notion that scientific explanation must be carried out in terms of our folk scientific conception of ourselves. A "folk theory" is just another name for our commonsense notions about a particular domain. For example, Aristotlean physics might be considered an explication of ancient Greek folk physics, a physics in which rocks fall because they desire to return to the place of their origin, and where heavier objects fall faster than lighter ones. Folk *psychology* would consist of the myriad rules of behavior humans use in their everyday relations with one another. (See Churchland[2] for a sketch of these rules.) Central to this folk theory is the liberal attributions of "beliefs," "desires," "moods," etc. to the entities that make up the domain of psychology: other people, pets, fictional characters, etc. In folk theories the issue is *not* whether they are useful abstractions or whether they are important to our day-to-day dealings with the world. (They are essential. Just reflect on the central role that folk psychological attribution plays in our justice system.) The issue is whether these common sense theories have any special status within science. In the case of contemporary scientific physics, it is accepted that folk physics has no special status. If physicists can explain the motion of bodies without anthropomorphizing them, then physics should do so.

The status of folk *psychology* is a different can of worms. As mentioned above, Paul Churchland has argued that not only *can* folk psychology be banished from a

mature scientific psychology, but the time has come to actually do so. In making his case against folk psychology, he mounts a three-pronged attack. First, he reminds us that we should not only assess a theory on its successes but also on its failings. And there is a large inventory of presumably psychological phenomena that folk psychology simply fails to address adequately, including the nature and dynamics of mental illness, creative imagination, sleep, perceptual illusions, and learning, just to name a few. Second, he argues that the history of folk psychology does not give one reason to hope for the future of the endeavor. Churchland writes that "the story [of folk psychology] is one of retreat, infertility, and decadence." It is a paradigm case of a degenerating research programme. Finally, Churchland outlines reasons for believing that folk psychology cannot be integrated easily with the rest of scientific explanations. Particularly, it seems to be very much at odds with the one field that it would presumably have the closest associations: neuroscience. On the weight of all three of these deficits, Churchland argues that the days of folk psychology in scientific psychology are numbered.

However, Churchland's EM in psychology is not without its objectors. Indeed, it is probably safe to say that it is still a minority view amongst philosophers of psychology. Some, like Dan Dennett[6,7] and Terence Horgan and James Woodward,[14] have argued that folk notions such as "belief" and "desire" should or must play a role in our scientific psychological explanations. For years, Dennett has argued the importance of an "intentional system" to psychological explanations. An intentional system is one which is "reliably and voluminously predicted" via the attribution of "beliefs," "desires," and other common sense notions to that system. And Dennett argues cogently that humans and other animals are just such systems. This being the case, a scientific psychology must employ concepts from folk psychology.

Horgan and Woodward take a slightly different approach. They argue that the case against folk psychology is overstated, that folk psychology is actually quite a good scientific explanation of psychology, regardless of its perported failings. In any case, they also argue that EM places too stringent restrictions on how folk psychology should be integrated with our other scientific beliefs. That neuroscience cannot capture the basic notions of folk psychology in its theory is no reason to reject it in favor of neuroscience.

But for all this heated debate over the importance of folk theory to psychology, we do not find anything even vaguely similar to biological theory. It is not clear whether such a debate is possible. The primary problem is determining whether a folk theory of biology exists in the first place. And, if a "folk biology" can be rounded up for the purpose, will its fate be more like that of folk psychology or folk physics?

One might look for a folk biology in the lore of the "common person," that general framework of common sense and rules-of-thumb that has served our species so well through the ages. Aside from common sense *psychological* knowledge about natural phenomena (e.g., "Always avoid contact with female bears when they are with their cubs, as mama bears are prone to protective violence when they *believe* their young are threatened"; "My dog is standing next to the door because he

wants to go out"; etc.), there seems to be little of what might be called specifically *biological* knowledge.

There is a good deal of folk knowledge of *breeding*, such as the old maxim that "like breeds like." The dangers of inbreeding and the knowledge that like animals will only mate with like animals have apparently been well known to breeders for centuries. Our first candidate for a folk biology, then, would be some version of the science of breeding. Indeed, part of the inspiration for Charles Darwin's *Origin of Species*[5] was the great diversity of pigeons that breeders had raised (even without knowledge of Mendelian genetics).

It is appropriate that Darwin's ground-breaking work should be mentioned, as its title names what arguably may be the central notion of any possible folk biology: the concept of a "species." The notion that the biological world is made up of distinct kinds of creatures is probably the first principle of common sense biology. The *Old Testament*, Native American mythology, and many other creation stories share the common feature that distinct kinds of creatures were created separately. Perhaps the biggest job of a scientific biology, from Aristotle onwards, has been the Herculean task of simply cataloging all the kinds of creatures found in our incredibly diverse ecosystem. The notion of distinct species is so central to our notion of what biology should be that this was what Darwin felt he had to explain with his theory of natural selection.

Along with this notion of diversity in the biological realm, perhaps another central notion to folk biology would be that the biological world constitutes a fundamentally different set of things, i.e., that there is something distinct and special about biological entities that separates them from the rest of the furniture of the universe. This notion of an essential difference between living and nonliving things is perhaps best captured in concept of the "vital spirit," the substance that is the essence of the living. Possession of this spirit makes a truly living cell different from a nonliving collection of the same chemicals. Though the popularity of the belief in some kind of nonmaterial animating "spirit" has suffered in this century, the spirit of the issue survives in the demands that society places on biologists and medical doctors to come up with reliable criteria of "life" and "death."

We are beginning to see that at least it is *possible* that something answers to the name "folk biology." It would have an ontology (that the world consists of the "biological" and the "nonbiological," and that the biological world is made up of distinct kinds or "species"). It would also have rules for the behavior between the elements of this ontology (like the laws of breeding). Folk biology might not seem to have the richness typically attributed to folk psychology (most of the breeding rules would seem to delineate all the things with which a given species *cannot* breed), but that might be because I simply have not adequately characterized it here. But one can imagine that a likely story might be put together. For the sake of argument, let us assume that such a likely story could be generated.

Even if the existence of folk biology is granted, it must be noted that, unlike the situation in psychology, there does not seem to be anybody interested in arguing for folk biology as the necessary or appropriate language of biological explanation.

Where there is vociferous debate in the philosophy of psychology, there is only silence in the philosophy of biology.

If the proceeding discussion has any cogency, it indicates that the current state of biology on the issue of eliminative materialism and the role of folk theory is different from that of psychology. This, in turn, indicates an area of disanalogy within the Sober analogy. However, this is not the most striking difference between the study of the mind and the study of life.

4.3 LACK OF THE SUBJECTIVE IN BIOLOGY

Here we come to the most striking difference between psychology and biology, a difference that probably underlies many of the other differences I have already sketched above. Psychological explanation has to explain *more* than just the behavior of psychological systems. One of the things that makes psychology such a difficult endeavor is that in addition to the straightforward *behavioral*, third-person phenomena which stand in need of explanation, in the case of humans at least, there seem to be additional *experiential*, first-person phenomena. Part of the burden of psychology is to explain (or explain away) phenomena related to the *prima facie* claim that psychological systems exhibit attention, intentionality, consciousness, self-consciousness, a "point of view," or the property of being "something-it-is-like-to-be" that entity, qualia, or any other of the constellation of concepts relating to the subjective nature of the psychological. Indeed, it seems plausible that this element of the psychological is what makes it so resistent to mechanistic or reductionistic explanation. It is the difficulty of even conceiving of a *conscious mechanism* that hampers the would-be psychological mechanist. Whatever consciousness is, apparently no collection of third-person facts about it would ever be complete; after science has done its best, there will still remain first-person facts inaccessible to the traditional scientific method.

It is not our place here to assess or take sides in the role or nature of consciousness in psychology. It needs only to be noted that, like the debate over eliminative materialism and folk psychology, there is no analogous concern in biology. Perhaps we should be thankful, for this is one less obstacle for theoretical biology to overcome or for AL to worry about. Biological phenomena, unlike their psychological counterparts, seem to be exclusively of the behavioral, third-person variety. There is no worry that, after describing all the physical parameters of the system, there still will be "something else." Now, determining what the correct parameters actually are and understanding exactly how biological systems produce the relevant behavior is a tough enough job on its own, but at least the phenomena in question are *there*—waiting to be measured, probed, and replicated. (I should note that Stevan Harnad makes many of these same points, more eloquently than I, in his contribution to these proceedings.[13])

To summarize the discussion so far, Sober proposed that the relationship between AL and biology was analogous to that between AI and psychology. This seems

to be a prominent point of view within the AL community. In fact, there seems to be support for the even stronger claim that the philosophy of AL should be the philosophy of AI translated into biological terms, a strategy I call the "global replacement strategy." However, in the last several pages, we have seen that a variety of issues and debates endemic to the philosophy of AI—those relating to eliminative materialism and the subjective nature of mind—seem to have no counterpart in biology. These issues cannot be discarded as being minor side issues within the philosophy of AI. Quite to the contrary, if the amount of ink spilled over them is any indication, they are among the most central philosophical issues of that endeavor. But, if these important issues cannot be translated into the philosophy of AL, what does this indicate about the general usefulness of GRS? It indicates that whatever the alleged validity and usefulness of translating concepts, problems, and metaphors from AI into AL as a constructive strategy, the GRS is clearly too extreme. For all the similarities between AI and AI, as indicated by the Sober analogy, the phenomena of intelligence and life are sufficiently different to preclude any kind of straightforward relationship between the two sciences.

5. CONCLUSION

In this paper, I have tried, through a variety of means, to suggest that a relationship between artificial intelligence and artificial life is not as useful as it might seem at first. Until this point in time, the philosophical discussion within AL has been littered with references to positions, metaphors, and arguments made popular within the history of AI. However, with the notable exception of Sober's 1991 paper, we have seen little discussion specifically of the methodology of importing concepts from AI into AL. By and large, the justification for this procedure simply has been accepted on the basis of the close intellectual ties between the two fields and their respective practitioners. The spirit of this paper is not that of a refutation of this methodology, but as a caution against its unreflective overuse.

We should not be surprised if concepts from AI are useful to AL. In this paper, I mention that Clark's concept of "microfunctionalism" is just such a useful construction. This particular example is not surprising in that Clark uses microfunctionalism in conjunction with his arguments for "connectionist" AI, an approach to AI that is arguably more *biological* than traditional approaches. It is the biological motivation of this position that leads me to suggest its usefulness to AL. Contrary to the GRS, it is not *necessarily* useful to AL because it is a concept from AI. One needs to make an argument for its usefullness beyond that provided by the Sober analogy. Such a burden of proof is laid upon anyone wishing to use any concept from AI in AL. The Sober analogy merely indicates a relationship between the two disciplines, and one should expect a *sharing* of ideas between them, not an eclipse of one by the other.

ACKNOWLEDGMENTS

I would like to thank Aaron Sloman, Marcus Peschl, Derek Smith, Tom Ray, Ron Chrisley, and Inman Harvey for enlightening discussion on the topics of this paper. Georg Schwarz and Sandra Mitchell both read drafts and, in the process of disagreeing with most of what I had to say, offered valuable criticism. I also would like to thank the members of UCSD's Experimental Philosophy Lab, for listening to me present this material in many forms. Earlier versions of this paper were presented at the 1992 Comparative Approaches to Cognitive Science Summer School (Aix-en-Provence), the University of Birmingham School of Computing Science, and at Artificial Life III, itself.

REFERENCES

1. Bedau, M. A., and N. H. Packard. "Measurement of Evolutionary Activity, Teleology, and Life." In *Artificial Life II* , edited by C. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X, 431–461. Redwood City, CA: Addison-Wesley, 1991.
2. Churchland, P. M. *Scientific Realism and the Plasticity of Mind.* Cambridge Studies in Philosophy. Cambridge, MA: MIT Press, 1979.
3. Churchland, P. M. *A Neurocomputational Perspective: The Nature of Mind and Structure of Science.* Cambridge, MA: MIT Press, 1989.
4. Clark, A. *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* Cambridge, MA: MIT Press, 1989: 34–36.
5. Darwin, C. *On the Origin of Species*, 1st ed. 1859.
6. Dennett, D. C. *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge, MA: MIT Press, 1978.
7. Dennett, D. C. *The Intentional Stance.* Cambridge, MA: MIT Press, 1987.
8. Farmer, D. F., and A. d'A. Belin. "Artificial Life: The Coming Evolution." In *Artificial Life*, edited by C. G. Langton. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. VI, 815–840. Redwood City, CA: Addison-Wesley, 1989.
9. Feyerabend, P. "Explanation, Reduction and Empiricism." In *Scientific Explanation, Space and Time*, edited by H. Feigl and G. Maxwell. Minnesota Studies in the Philosophy of Science, Vol. 3, 28–97, 1962.
10. Feyerabend, P. "Materialism and the Mind-Body Problem." *Rev. Metaphys.* **17** (1963): 49–66.
11. Feyerabend, P. "Mental Events and the Brain." *J. Phil.* **60** (1963): 295–296.

12. Harnad, S. "The Symbol Grounding Problem." *Physica D* **42** (1990): 335–346.
13. Harnad, S. "Artificial Life: Synthetic vs. Virtual." This volume.
14. Horgan, T., and J. Woodward. "Folk Psychology is Here to Stay." *Phil. Rev.* **XCIV** (April 1985): 197–226.
15. Laing, R. "Artificial Organisms: History, Problems, Directions." In *Artificial Life*, edited by C. G. Langton. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. VI, 49–61. Redwood City, CA: Addison-Wesley, 1989.
16. Langton, C. G. "Artificial Life." In *Artificial Life*, edited by C. G. Langton. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. VI, 1–47. Redwood City, CA: Addison-Wesley, 1989.
17. Pattee, H. H. "Simulations, Realizations, and Theories of Life." In *Artificial Life*, edited by C. G. Langton. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. VI, 63–77. Redwood City, CA: Addison-Wesley, 1989.
18. Ray, T. "An Approach to the Synthesis of Life." In *Artificial Life II*, edited by C. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X, 371–408. Redwood City, CA: Addison-Wesley, 1991.
19. Searle, J. R. "Minds, Brains, and Programs." *Behav. & Brain Sci.* **3** (1980): 417–424.
20. Searle, J. R. "Consciousness and Cognition." *Behav. & Brain Sci.* **13** (1990).
21. Sober, E. "Learning from Functionalism: Prospects for Strong Artificial Life." In *Artificial Life II*, edited by C. Langton, C. Taylor, J.D. Farmer, and S. Rasmussen. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X, 749–765. Redwood City, CA: Addison-Wesley, 1991.
22. Stich, S. "From Folk Psychology to Cognitive Science." In *The Case Against Belief*. Cambridge, MA: MIT Press, 1983.
23. Turing, A. M. "Computing Machinery and Intelligence." In *Minds and Machines*, edited by A. Anderson. Engelwood Cliffs, NJ: Prentice Hall, 1964.