What's Life Got To Do With It?[1]

Margaret A. Boden [2]

Abstract.

This is a re-formatted version of section 16.x of M. A. Boden, *Mind As Machine: A history of Cognitive Science (Vols 1–2),* Oxford University Press, Oxford, 2006, It discusses the question wether mind can occur without life.

The book has a large bibliography. All the references in this extract are to items listed in the bibliography.

It is included in the AIIB proceedings and web page with the kind permission of the author. Letters in the section headings refer to subsections in the book.

1 Introduction

All the minds we know about are found in living things. But why?

- Couldn't there be mind and meaning without life?
- And what is life?
- Given that it involves self-organization, just what sort of self-organization is it?
- Must it involve evolution, for instance?
- Is embodiment essential? And what's that? Is mere physicality enough for embodiment?
- Or is metabolism needed too–and, again, what's that?
- What's the link, if any, between metabolism and mind?
- Could life be generated artificially, and if not why not?
- Is strong (i.e. virtual) A-Life impossible in principle?
- If so, does it follow that strong AI is an illusion too?

These questions took a long time to surface in cognitive science. Or perhaps one should rather say to *resurface.* For in the very early days, life and mind were both discussed–and were treated largely on a par.

The cyberneticians of the 1940s applied their theories of self-controlling machines to both living organization and purpose (see 4.v-vii). Life wasn't regarded as necessary for teleology, for self-guided missiles were said to exemplify goal seeking (Rosenblueth, Wiener and Bigelow 1943). However, since life and mind were supposed to consist in fundamentally similar principles of control, it didn't seem surprising that all the minds we know about are grounded in living things.

McCulloch had been a member of this life-and-mind movement since its inception in the 1930s. But he'd been deeply interested in logical analyses of language even before then (see 4.iii.c). In the event, his paper of 1943 turned attention away from life in favour of mind.

Control by feedback gave way to logic-based computation. The NewFAI computer-modelling community saw mind and meaning as matters for propositional logic, their origin in adaptive behaviour and living embodiment being downplayed. Even unreconstructed cyberneticians now sometimes spoke of mind without mentioning life: when William Ross Ashby wrote a paper on cybernetics for *Mind,* he concentrated on defending materialism against mind-body dualism, not on exploring the philosophy of self-organization (Ashby 1947).

In short, by mid-century the link between life and mind–though still widely accepted, even taken for granted–wasn't explicitly considered. The people on the NewFAI side of the emerging cybernetics/symbolic schism (4.ix) ignored life and spoke only of mind. The people on the other side said very little about mental phenomena beyond perception and goal-seeking: self-reflection and reasoning were mostly ignored.

That life/mind split within cognitive science lasted for several decades. Much the same was true in philosophy, especially the analytic variety. It's still the case, in 2005, that the link between life and mind is ignored (beyond mere lip-service) by the mainstream. However, these issues are now arousing interest–largely as a result of the rise of A-Life.

2 a: Life in the background

Even in the early 1960s, there were a few exceptions to what I've just said. Most important, with respect to their historical role in cognitive science, were the people impressed by cybernetics who were developing holistic philosophies–and computer models–of life and/or mind (15.vi-vii).

The most influential of these (all discussed later in this Section) were to be the Chilean neuroscientist Maturana (with Varela and Milan Zeleny), whose similarity to Dreyfus was mentioned in Section vii.b above, and the philosopher Howard Pattee, with his students Robert Rosen and Peter Cariani. Both Maturana and Pattee were working on the concept of life from the 1960s. Although they had some early disciples, their ideas didn't become widely known until the 1990s. By the new millennium, however, Maturana and Varela's ideas had been published in semi-popular form, and an entire issue of the journal *BioSystems* was devoted to Pattee's work and influence (Rocha 2001).

All of the above were either scientists (like Maturana) or philosophers very close to science, so close that they sometimes got involved in scientific work (Pattee, for instance). None were "pure" philosophers.

In general, the mid-century philosophers who were interested in the puzzle of life-and-mind came from the Continental, not the analytic, side of the fence. As a result, they were largely ignored by the

[2] University of Sussex, UK, M.A.Boden@sussex.ac.uk

scientific community. Moreover, they were exceptions even within their own, neo-Kantian, tradition. To be sure, phenomenologists in general took the human being's embodied living-in-the-world as philosophically basic. And Wittgenstein saw language as part of our "natural history", declaring: "What has to be accepted, the given, is– so one could say– *forms of life*" (1953: 226, italics in original). Most of them had scant interest, however, in what biologists mean by life– which includes oak trees and barnacles, as well as human beings.

One exception to this was the existentialist theologian Hans Jonas (1903-1993), who developed a new philosophy of biology in the 1950s. Unlike Maturana and Varela, he wasn't interested in biology for its own sake, but as an aspect of what he saw as the disastrous cultural *denouement* of Descartes' materialism (Jonas 1966: 58-63).

An ex-pupil of Heidegger, Jonas fled Germany for England when the German Association for the Blind expelled its Jewish members (Jonas 1966: xii). From the mid 1950s to the mid 1970s he worked at the New School of Social Research, in New York. Despite still regarding Heidegger as "the most profound and ... important [proponent] of existential philosophy" (229), he rejected his philosophical dichotomy between humans and other living things–and his pro-Nazi sympathies, too. He explained the latter in terms of "the absolute formalism of [Heidegger's] philosophy of decision," in which "not *for* what or *against* what one resolves oneself, but *that* one resolves oneself becomes the signature of authentic Dasein" (Jonas 1990: 200). And that, in turn, he saw as a result of the stripping-away of value from nature, its "spiritual denudation" by Descartes and modern science (Jonas 1966: 232).

It was in response to this disenchantment of nature that Jonas published various essays on life in the post-war years, and collected them as *The Phenomenon of Life* in 1966. They outlined a framework for a biology that would admit value as an intrinsic feature of life in general. ("Outlined" and "framework" are important here: he discussed almost no specific examples.)

Embodiment, and in particular metabolism, was seen by Jonas as philosophically crucial (1966: 64-91). Not only was life essential for the emergence of mind (99-107), but *all* self-organized matter was, in a sense, ensouled–though where Maturana and Varela spoke of life as involving *cognition,* Jonas spoke of life as involving *self-concern.* (He lauded Heidegger for having "shattered the entire quasi-optical model of a primarily *cognitive* consciousness, focusing instead on the wilful, striving, feeble, and mortal ego"–1996: 44; italics in original.) As he put it:

> One way of interpreting [the ascending scale of life] is in terms of scope and distinctness of experience, of rising degrees of world perception.... Another way, concurrent with the grades of perception, is in terms of progressive freedom of action.... [One] aspect of the ascending scale is that in its stages the "mirroring" of the world becomes ever more distinct and self-rewarding, beginning with the most obscure sensation somewhere on the lowest rungs of animality, even with the most elementary stimulation of organic irritability as such, in which somehow already otherness, world, and object are germinally "experienced," that is, made subjective, and responded to. [We spoke, above, of freedom.] One expects to encounter the term in the area of mind and will, and not before: but if mind is prefigured in the organic from the beginning, then freedom is. And indeed our contention is that *even metabolism, the basic level of all organic existence, exhibits it: that it is itself the first form of freedom* (Jonas 1966: 2f.; italics added).

Even in "the blind automatism of the chemistry carried on in the depths of our bodies," there is "a principle of freedom ... foreign to suns, planets, and atoms". For living organisms have a special type of identity and continuity: a stable dynamic form made of an ever-changing material substrate. In short, "mind is prefigured in organic existence as such" (5). Plants, too, have "metabolic needs", although they stand in an "immediate" relationship to their environment. And metabolism is the necessary base of all forms of mediation: perception, motility (action), emotion, and–ultimately–conscious imagination and self-reflection. (These phenomena emerge as a result of evolution: Darwin, *despite* his materialist assumptions, had enabled us to understand this: 38-58.) Life and mind are ontologically inseparable: "the organic even in its lowest forms prefigures mind, and ... mind even on its highest reaches remains part of the organic" (1).

In other words, Jonas was offering an answer to the question of *why* all the minds we know about are found in living things. At the same time, he was offering an answer to the question of *what life is.*

He explicitly refused to speculate about the origins of life (4), even though this was already being discussed by biochemists (Chapter 15.x.b). His interests were ontological, not scientific: metabolism was "the break-through of being" from mere physicality to "the indefinite range of possibilities which hence stretches to the farthest reach of subjective life ..." (3). (Accordingly, he retained a Heideggerian hostility to technological theories/analogies of life or mind: 108-126.)

The book was reissued (by several different publishers) in 1979, 1982, and 2001, and translated into German in 1994. So one can't say that it was wholly ignored. Indeed, because of his stress on the intrinsic value of life and humankind's responsibility towards it, Jonas' work–especially his volume on ethics (Jonas 1984)–has recently become better known thanks to the environmentalist movement.

In the 1950s and 1960s, however, his philosophy of biology was ignored by analytical philosophers and mainstream biologists alike. (And by cyberneticists too, whose analysis of living purpose and rocket teleology *in the very same terms* he'd rejected as "spurious and mainly verbal"–1966: 111.) The same was true of Maturana and Varela's early work, but they have now earned a clear, if still marginal, place in the history of cognitive science. Jonas has not (but see Di Paolo in press). He's relevant here not as a protagonist in that historical drama, but as a mid-century philosopher who tried to argue the case that mind requires life, rather than taking it for granted.

Another philosopher who'd done this was Henri Bergson (Chapter 2.vii.c). By the end of the twentieth century, Bergson's views on "creative evolution" were being revived in some philosophical circles–especially in "process" philosophy/theology (Sibley and Gunter 1978; Papanicolaou and Gunter 1987). This emulated Bergson alongside the even greater hero Alfred North Whitehead (4.iii.b). But some unorthodox scientists were taking an interest too. The physicist/philosopher Henri Bortoft (1996) put Bergson second only to Goethe as a precursor of current dynamical theories in science and philosophy (see 2.vii.c). And a few neo-Bergsonian philosophers even tried to relate his ideas to cognitive science and/or A-Life.

For example, Gilles Deleuze (1925–) revived certain aspects of Bergson's philosophy by stating them in terms of ideas about dynamical systems (Deleuze 1966/1988). I'm saying that at second hand, I must confess, for Deleuze himself is nigh unreadable by anyone more accustomed to analytical philosophy. Much as Richard Montague's work couldn't spread among linguists until a clear account of it had been provided by Barbara Partee (see 9.ix.c), so Deleuze's has been

made accessible to cognitive scientists by his expositor Manuel DeLanda (2002).

Although he rejected Bergson's dualist interpretation of *elan vital,* Deleuze offered a "re-enchantment" of matter, nevertheless. He even (confusingly) used the term "spirituality" in talking about matter and life. But this wasn't intended as transcendent spirituality: rather, it referred to the abstract principles of self-organization, and the structured spaces of possibilities, that are inherent in matter/energy.

He saw matter not as inert stuff subject to external influences, but as the source of formative material *processes.* A soap-bubble, for instance, actively minimizes the surface tension at every point (it dynamically "computes" its own shape). The dynamical structures generated by matter were said to be constrained, in part, by abstract topological principles describing connectivities and attractors of various kinds (compare Stuart Kauffman's work on NK networks, and Randall Beer's on CTRNs: 15.ix.b and xi.b).

On this view, life was a special case of matter, and mind a special case of life. It followed that there's no *special* difficulty about giving a naturalistic, even a materialistic, account of mind or intentionality, even though spelling one out in detail may be highly challenging.

However, these intriguing analogies weren't helpful in furthering scientific understanding. (Or anyway, they haven't been helpful yet: DeLanda's relatively accessible version of Deleuze appeared only two years ago, and it remains to be seen whether many scientists will take it up.) Admittedly, the ever-maverick neuroscientist Karl Pribram (1987)–accused in the early 1960s of actually *believing* the MGP manifesto (6.iv.c)–described the cerebral basis of some cognitive processes in Bergsonian terms. But that's not to say that he *used* Bergson's ideas to make discoveries which otherwise would not have been made. Rather, he pointed out an analogy between Bergson's views on memory and his own (longstanding) holographic/holonomic theory (cf. 12.v.c).

Cognitive scientists who weren't already sympathetic to dynamical systems and/or Kauffman's approach to A-Life weren't likely to be interested in Bergson's work at all, even if they encountered it. And that was unlikely: as remarked in Chapter 2.vii.c, it had been more or less forgotten since mid-century–especially by philosophers of an analytic cast of mind.

For over thirty years, then, the concept of life was usually ignored in discussions of mind as machine. To be sure, the psychologist Miller raised the topic–but he immediately dropped it like a hot potato. He was, he said, "unclear" whether epistemic (cognitive) systems should be defined as animate or inanimate. The advantage of defining them as animate was that "we cut artificial intelligence free to develop in its own way, independent of the solutions that organic evolution happens to have produced" (Miller 1978: 9). (This remark predated the concept of strong A-Life by a decade: clearly, Miller thought it obvious that computers and life are incompatible.) But whether it made "any real difference" in conceptualizing the study of *mental* processes was "unclear".

Most analytic philosophers tacitly assumed some life-mind linkage–which would imply that if computers aren't alive then they aren't psychological systems. They evidently thought this point so obvious that, even when they bothered to state it explicitly, they didn't offer any arguments for it.

Scriven, for instance, confidently declared–without giving reasons–that "Life is itself a necessary condition of consciousness" and that "Robots ... are composed only of mechanical and electrical parts, and cannot be alive" (Scriven 1953: 233). Lucas hinted at a similar position in his own reply to Turing's 1950 paper (see Section v.a). Geach insisted that AI systems can't have beliefs and intentions because they're "certainly not alive" (Geach 1980: 81). And some, such as Searle (1980, 1992) and Ruth Millikan (1984), explicitly linked intentionality with biology (neurochemistry and evolution, respectively). But even they didn't discuss the nature of life as such.

Two exceptions that proved (i.e. tested) the rule were Putnam's (1964) paper on "Robots: Machines or Artificially Created Life?" and Geoffrey Simons' (1983) book *Are Computers Alive?*

Despite its title, Putnam's paper focussed mainly not on life, but on consciousness. At one point, Putnam endorsed Ziff's claim that it's an "undoubted fact" that if a robot isn't alive then it can't be conscious. But he was relying on "the semantical rules of our language", not on any quasi-explanatory relationship between life and mind.

He also said (this time, disagreeing with Ziff) that something which is clearly a mechanism might be alive. Again, however, this was linguistic philosophy in action. Sometimes, Putnam heretically recommended *changes in meaning* due to new scientific data, as he did when countering Malcolm's account of dreaming (Putnam 1962). But in the paper on robots and life, he was talking only about what *current usage* allowed one to say (or imagine) without contradiction. The nearest he got to discussing a substantive claim about life was to scorn the suggestion that the primary difference between a robot and a living organism is the "softness" or "hardness" of the body parts (1964: 691).

Much later, Putnam's paper was discussed at length, and accused of incoherently combining Aristotelian and Cartesian views (Matthews 1977). At the time, however, it didn't prompt philosophical interest in the concept of life.

Simons, writing twenty years after Putnam, used concepts drawn from GOFAI and cybernetics to claim that computers can be *really* alive, and *really* intelligent. He specifically denied that the genesis of the system is relevant to whether it's alive: "A mechanically *assembled* [i.e. not evolved or self-constructed: see below] system may reasonably be regarded as living .... " (1983: 23). However, his argument was neither deep nor convincing, and (deservedly) attracted little attention.

3 b: Functionalist approaches to life

With the rise of A-Life in the late 1980s, the nature of life became an inevitable topic for computational research. Inevitable, but in practice not central: most A-Life workers focussed on other questions, maintaining a diplomatic silence on this one. Some of their colleagues, however, were more bold.

The relevant discussions were guided by two radically opposed philosophies. (Sounds familiar?–see Section vi.b.) These were functionalism and metabolic holism, a special case of dynamical systems theory.

Functionalism, in this context, is the view that the characteristics of life (see Chapter 15.ii.b) can be described by informational concepts. So self-organization involves the appearance of new levels of order, abstractly defined. Autonomy, emergence, development, adaptation, responsiveness, and evolution concern various types of structure, process, and control. Even reproduction (on this view) can be defined informationally, as self-copying.

The one exception is the concept of metabolism, which concerns not information but energy. Thoroughgoing A-Life functionalists weren't worried by this, as we'll see. But their opponents argued that they should be.

It's often assumed (wrongly) that all A-Life workers are thoroughgoing functionalists. This is largely because Christopher Langton, following John von Neumann's lead, wrote this position into his definition of the field in 1986 (Chapter 15.ii.b and ix).

Moreover, he drew the obvious implication: a licence for strong A-Life. If living self-organization is definable in logical terms, then a virtual "creature" implemented in computer memory that satisfied these abstract criteria–whatever they are–would be genuinely alive. ("Whatever they are", because definitions differed. For instance, Langton suggested including the lambda parameter, Andrew Wuensche the Z-parameter: Chapter 15.viii.a.)

Some A-Life colleagues were quick to join Langton in this claim. Thomas Ray, for instance, declared:

> The intent of [my] work is to synthesize rather than simulate life .... To state such a goal leads to semantic problems, because life must be defined in a way that does not restrict it to carbon-based forms. It is unlikely that there could be general agreement on such a definition .... Therefore, I shall simply state my conception of life in its most general sense. I would consider a system to be living if it is self-replicating, and capable of open-ended evolution [generating] structures and processes that were not designed-in or preconceived by the creator (Ray 1992: 372).

As we saw in Chapter 15.vi.b, Ray's *Tierra* system did indeed generate phenomena not designed-in by Ray. These included co-evolving parasites, hyper-parasites, cheaters, and symbionts. Ray's response was a curious combination of modesty and hubris:

> [The] results presented here are based on evolution of the first creature that I designed, written in the first instruction set that I designed. Comparison with the [virtual] creatures that have evolved shows that the one I designed is not a particularly clever one .... It would appear then that it is rather easy to create life (p. 393).

As for the problematic concept of metabolism, Ray said two things. On the one hand, the computer consumes physical energy too. On the other, the equivalent of metabolism can be functionally defined:

> In studying the natural history of synthetic organisms, it is important to recognize that they have a distinct biology due to their non-organic nature. In order to fully appreciate their biology, one must understand the stuff of which they are made.To study the biology of creatures of the RNA world would require an understanding of organic chemistry and the properties of macro-molecules. To understand the biology of digital organisms requires a knowledge of the properties of machine instructions and machine language algorithms (p. 397).
>
> I will discuss the inoculation of evolution by natural selection into the medium of the digital computer. This is not a physical/chemical medium; it is a logical/informational medium .... Evolution is then allowed to find the natural forms of living organisms in the artificial medium. These are not models of life, but independent instances of life (Ray 1994: 179).

For some broadly functionalist A-Life scientists, this was a step too far–and *much* too far for most philosophers (e.g. Harnad 1994; Olson 1997). Those A-Life colleagues were content to interpret most of the characteristics of life in informational terms–but not metabolism, which is irredeemably physical. However, since they defined metabolism as mere energy dependency, their rejection of Ray's position was intuitive rather than strongly argued (see below).

Many A-Life colleagues simply avoided the question, by way of the "diplomatic silence" mentioned above. They were interested in studying specific aspects of life, such as evolution or flocking, not in discussing its general nature. They were even less interested in considering the "strong A-Life" scenarios sketched by Langton and Ray–and later by Steve Grand (1958–).

Grand's first claim to fame was that he designed the hugely popular computer game *Creatures.* This swept the world in the early 1990s (see Chapter 13.vi.d), and was still being widely celebrated–for broadly counter-cultural reasons–in the new century (Kember 2003: 91-105).

Creatures enabled the user to evolve unusually sophisticated computer creatures (Grand and Cliff 1998). Their neural-network brains supported simple learning, and included 'neuromodulators' as well as several types of 'neurone.' The creatures also had a simulated biochemistry, with the potential to model a large number of metabolic and hormonal functions, from digestion to ovulation. As a piece of life-like software engineering, it was way beyond the general state of the art when it appeared, and is still impressive. Indeed, it could conceivably be used as a powerful testbed for AI models of motivation and emotion such as those discussed in Chapter 7.i.e-f (Boden 2000b).

Grand's current technical aim is to build an "imaginative" robot called Lucy, whose intelligence will emerge "naturally"–and holistically–from its 100,000-neurone hardware (Whitby and Grand 2001). As he points out, this attempt to build Dennett's (1978c) "whole iguana" is very different from MIT's Cog project, with which Dennett himself was involved (see 15.vii.a).

The Cog robot was carefully designed module by module, bits of its "intelligence" being successively bolted on. Grand, by contrast, wants an already integrated intelligence to emerge from a relatively unorganized base. Rather than providing Lucy's brain with spatial maps or orientation columns, for instance, he hopes that these would emerge spontaneously (much as ocular dominance columns arose in the work of Christoph von der Malsburg and Ralph Linsker: see Chapter 14.vi.b and ix.a). And the robot would learn to perform "voluntary" actions by associating the image (representation, model) of the desired action with the muscle movements required to achieve it (compare Marr's theory of the cerebellum: 14.v.c).

The Lucy project is startlingly ambitious–I'm tempted to say, utterly impracticable. But the A-Life expert David Cliff (p.c.) believed *Creatures* to be utterly impracticable too, when first consulted by the games company to whom Grand had offered it. Given what Grand had told them it could do, it must–so Cliff thought–be either hype and/or a superficial con-trick, carefully tailored to present a convincing 'demonstration.' (Even the impressive SHRDLU, you'll remember, could handle only the one conversation without tripping over its toes: 9.xi.b.) And the fact that it had been two-finger-typed on Grand's bedroom computer wasn't promising. Not until he got down into the machine code was Cliff convinced–at which point he suggested how it could be improved still further, using some of the ideas discussed in Chapters 7.i.f and 15.vi-ix.

Grand didn't know about those ideas already, because he's an autodidact. As such, he's undeterred by received academic opinion. And he's a highly creative computer engineer, who's already designed one apparently impossible system that does just what it was intended to do. He's thus in an entirely different class from the self-publicizing roboticists Kevin Warwick and Hugo de Garis, on whose 'research'–technical no less than philosophical–I forbear to comment, for fear of scorching the page.

I wouldn't bet a large sum of money on Lucy. And I don't agree with those cultural commentators who claim that Grand is "one of the 18 scientists most likely to revolutionise our lives in the coming century" (n.a. 2000b). Nevertheless, as Richard Dawkins has remarked, "If anybody can pull off a spectacular breakthrough, it'll probably be him" (Whitby and Grand 2001: 13). (For the most recent status-report on Grand's progress, see his website at http://cyberlife-research.com/people/steve/.)

At the turn of the century, Grand (2000, 2003) made a number of highly provocative claims about the philosophical significance of his own past and future work. He sees his virtual creatures as more than merely *life-like:* they are "sort of alive", or even "a sort of life". When challenged on this point, he insists (p.c.). Grand is an autodidact in philosophy too, but here there's no good reason to give him the benefit of the doubt. Whereas *Creatures* (considered as technology) clearly does what he said it would do, his philosophical arguments are challengeable–and, in my view, as mistaken as Ray's. Strong A-Life is no more plausible in *Creatures* than in *Tierra* –and even Grand's predicted robot Lucy wouldn't count as genuinely *alive* (see the discussion of metabolism, below).

Where the general public were concerned, Lucy made something of a splash. Although it must be said–and often is said, by other roboticists–that if Grand hadn't fitted a furry gorilla-face onto the head, and if he'd called it "Robot 37" instead of "Lucy", people wouldn't have been quite so interested. (Similarly, the young Minsky's robot arm aroused no attention until he put a shirtsleeve on it: see 1.iii.h.) Quite apart from the over-excitement of the journalists (the same old story!), a number of commentators have picked up on it as an expression of wider cultural concerns.

The anthropologist Lucy (sic!) Suchman, for example, who cast doubt on GOFAI planning some twenty years ago and focussed on *human-machine communication* soon after that (13.iii.b), described her robotic namesake as one among the disturbing category of the "almost human" (Castaneda and Suchman forthcoming; cf. Suchman 2004).

Besides the familiar anthropologists' fare of totems and other things "doing duty as persons", these include children, non-human primates, and AI/A-Life machines. The cultural status of children has been a focus of commentary at least since Jean-Jacques Rousseau (1712-1778), and twentieth-century developmental psychology has helped fuel this fire. As for primates, advances in field ethology have led to the culturally problematic Great Ape programme (7.vi.f). The eighteenth-century automata (2.i.b) challenged contemporary notions of the person (Riskin 2003). Now, as Suchman pointed out, various actual and imaginary AI projects are exciting comment not only in the philosophy of mind but in our wider culture too.

Lucy (which Suchman discusses at length) is only one example of the "almost human" produced by AI/A-Life. Cog, and especially its successor Kismet (see 13.vi.d), are others. The feminist philosopher Evelyn Fox Keller (forthcoming), for instance, sees some "serious anxieties" with respect to providing Kismet and the like with facial expressions that reliably elicit emotional reactions in human viewers. (She's particularly worried by the plan to use robots like Brian Scasselatti's Nico to *test* theories in human developmental psychology.) Still other almost-humans–all media darlings in their day–include ELIZA, expert systems, AI agents ("softbots"), VR avatars, Turing's computer conversationalist, Stanley Kubrick's HAL, and Stephen Spielberg's David.

The behaviour–and man-machine interactions–of many of these systems is far more humanlike than Lucy's is. But because Grand, besides providing the superficial furry face, speaks of *life* as well as *mind,* his work aroused more outside comment than most. In addition, his A-Life system is not virtual/intellectual (as softbots are) but *embodied*– or at least, *material.* It's therefore of interest to those commentators, including phenomenologists and many feminist philosophers, for whom the downplaying of embodiment in the analytic/scientific tradition has been a fundamental mistake (Haraway 1997; 186, 302f. and passim; Kember 2003: 105-115, 198ff., and passim).

Suchman (like Haraway, and also Clark: vii.d, above) takes personhood, in whatever culture, to be constituted not by an individual person-in-the-mind but by the nexus of social relations and interactions available. On that view, the cultural status of robots, and other AI/A-Life systems, is determined less by their seeming intelligence than by the pattern of interactions we choose to engage in with them. But the influence is reciprocal: insofar as we do engage with them, we modify our own self-image in various subtle ways (cf. 13.vi.d).

4 c: The philosophy of autopoiesis

Some A-Life researchers dismissed all these science-fictional scenarios *because they were fundamentally opposed to functionalism in the first place.* Among these were the proponents of Maturana's theory of "autopoiesis".

This was perhaps the best-developed philosophy of metabolic holism. (The competing candidate is the work of Pattee's group: see below.) It even inspired several computer models of biochemical autopoiesis (Zeleny 1977; Zeleny, Klir and Hufford 1989), and a wide range of work in A-Life (McMullin 2004). This included work in "wet" A-Life, in which biochemical autopoiesis *as such* was studied too (see Chapter 15.x.b). (Bachman et al. 1990; Walde et al. 1994).

Originated in the 1960s, Maturana's theory was strongly influenced by Heinz von Foerster's cybernetics. Despite the fact that an English translation was published over a quarter-century ago in the highly respected "Boston Studies in the Philosophy of Science" (Maturana and Varela 1972/1980), it has remained a minority taste. It's clear from my personal acquaintance that many philosophers have never even heard of it.

One reason, no doubt, is its rebarbative vocabulary and unrelenting abstraction. Also, it has some highly counterintuitive implications, as we'll see. Nevertheless, it offers a principled way of grounding mind in life. Rather than arguing (like Searle) that neuroprotein happens to cause intentionality, as chlorophyll happens to cause photosynthesis, this view grounds intentional categories in an essentially autopoietic biology.

For Maturana and Varela (1972/80), life is "autopoiesis in the physical space". Autopoiesis in general is defined as the continuous self production of an autonomous entity. As they put it (you're advized to take a deep breath here):

> An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components which: (i) through their interactions and transformations continuously regenerate the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network (Maturana and Varela 1972/80: 79).

Or more colloquially, an autopoietic system "pulls itself up by its own bootstraps and becomes distinct from its environment through its own dynamics, in such a way that both things are inseparable". This type of self-organization can occur in the world of human communication, in which case we have some kind of social institution (cf. Teubner 1987, 1993). But when it happens in the physical world, we have a living organism.

The autopoiesis concerned here is a special case of homeostasis (see Chapter 4.v.c), where what's preserved isn't one feature, such as blood temperature, but the organization of the system as a unitary whole. This requires the self-creation of a unitary physical system, by the spontaneous formation of a boundary–at base, the cell membrane–and the continuous generation and maintenance of the body's own components.

For Maturana and Varela, *body* and *embodiment* are autopoietic categories. So too are *cognition, communication, meaning,* and *language,* all of which they defined in terms of the interactions of living things. In the more accessible version of their theory that appeared around 1990 (Maturana and Varela 1987, 1992), and in Varela's book coauthored with cognitive psychologists (Varela, Thompson and Rosch 1991), they focussed on human language, understanding, society, and consciousness–all described as necessarily rooted in our biology.

In fact, they were overly liberal with their ascriptions of intentionality (Boden 2001), for they declared that "Living systems are cognitive systems, and living as a process is a process of cognition" (1972/80: 13). Taken seriously, this extends knowledge even to algae and oak trees. One can–and should–express the idea that algae and acorns are pre-adapted to their environment without using the concept of *knowledge.* Such over-liberality was an occupational hazard for cyberneticians: as we saw in Chapter 4.v.e, Gregory Bateson had similarly attributed *knowledge* to redwood forests, and *mind* to whirlpools and oscillating electrical circuits.

From the autopoietic viewpoint, both strong A-Life and strong AI are absurdities. For computers aren't autopoietic systems. Even self-assembling robots, if assembled from manufactured parts as opposed to being self-organized by some alien biochemistry, wouldn't be alive. (Nor would they have bodies.) Consequently, robotic intelligence is impossible too.

Autopoietic theory is a special case of the general (anti-functionalist) position that metabolism is essential for life. Believers in strong A-Life (such as Ray, quoted above), when confronted with this view, typically pointed out that computers consume energy too. They sometimes added that the "physics and chemistry" of their virtual creatures is constituted by the computer's memory and operating system.

A number of philosophers, some of whom weren't committed to autopoietic theory, replied that metabolism is more than mere energy dependency. Rather, it's the self production and self maintenance of the physical body by energy budgeting, involving self-equilibrating energy exchanges of some *necessary* complexity (Pattee 1989; Cariani 1992; Sober 1992; Boden 1999). They argued that strong A-Life is possible only if virtual systems can metabolise *in the sense just given,* or if metabolism is inessential for life. But neither alternative is tenable.

Living 'tin-can' robots are also excluded by this approach. Only robots powered by complex biochemical cycles of synthesis and breakdown would be truly alive, and truly embodied. This is the basis of the intuition scorned by Putnam, that "softness" and "hardness" matter (see above).

Elliott Sober (1948–) cited other biological properties, besides metabolism, in arguing against strong A-Life (Sober 1992). Digestion and predation (for example) each relate an organism to something outside itself, where that "something" is essentially physical. Both can be realized in multiple ways (defined by biochemistry and behaviour), but in every instance some physical organism has to interact with–hunt, eat, transform–another. Like metabolism itself, these features can be usefully simulated by A-life models. But they can't be replicated, so strong A-Life is impossible.

Sober's argument would be endorsed by autopoietic theorists. But to see metabolism as essential for life isn't necessarily to accept autopoietic philosophy. For this has some surprising implications, which many people reject. One was noted above, namely, the conflation of life and cognition. Another was remarked in Chapter 15.viii.b: the embargo on terms such as *input, output, function, feature detector,* and *representatiom.* Two more concern features often listed in definitions of life: reproduction and evolution.

Maturana and Varela's claim that the formation of the cell membrane is *the* fundamental phenomenon of biology, and that life involves the "total subordination of [all the processes of change within] the system to the maintenance of its unity" (1972/80: 97), implied that reproduction isn't essential for life. For them, this process is not (as functionalists claim) informational self-copying, but the formation of new autopoietic unities from previous ones. It follows that life is prior to reproduction (pp. 105-107). This wasn't a merely conceptual point, but a substantive biological hypothesis: that the earliest living organisms needn't have been able to reproduce (Boden 2000b).

Evolution, also, was seen by them as inessential, because it requires reproduction. (Inessential for life, but not for what's normally regarded as knowledge: they admitted that only evolution can generate the complex organisms typically credited with cognition.)

This conclusion, though unusual, is less controversial. For, *pace* Ray, and many theoretical biologists too (e.g Maynard-Smith 1996), there are three independent arguments against defining life in terms of evolution. First, populations, not individual organisms, would be paradigm cases of life. Second, creationism would be conceptually incoherent, not just false. And third, a population in evolutionary equilibrium wouldn't count as alive.

5 d: Evolution, life, and mind

Some philosophers of A-Life, nevertheless, took evolution (together with metabolism) to be the sort of self-organization which characterizes life. Pattee was an early example, followed by his students Rosen (1985, 1991) and Cariani (1992, 1997). He'd modelled co-evolution in the 1960s (see Chapter 15.vi.a). Subsequently, he focussed on the emergence of new phenotypic structures and functions.

A crucial example, for Pattee (1985), was novel types of "measurement", or classification. These were understood as ranging from enzyme activity to sensory perception–as in the evolution of new sensory organs (see Chapters 4.v.e and 15.vi.d). Pattee's concept of measurement was intriguingly similar to Smith's "participatory registration"–but, unlike Smith, he retained the first definition of computation distinguished in Section ix.a. So he specifically dismissed strong A-Life, arguing that measurement requires physical interaction, which can't be realized by formal computational systems. *A fortiori,* no novel biological functions can emerge in formal evolutionary systems (15.vi.d). He did allow, however, that "weak" A-Life modelling (simulation) could help clarify central biological and psychological concepts.

In the 1990s, another philosopher of A-Life argued that evolution is an essential criterion. Mark Bedau (1954–) explicitly accepted the three counterintuitive implications mentioned above, because of the explanatory power gained by defining life in evolutionary terms (Bedau 1996). And this explanatory potential, he said, was augmented by A-Life. In presenting his account of "supple adaptation" (alias evolution), he argued that A-Life modelling can deepen our understanding of life as such, because it helps us to study evolution in dynamic and quantitative terms. Moreover, he extended his evolutionary argument from life to mind (Bedau 1999, in prepn.).

A-Life philosophers weren't the only ones to link life and mind. Others, too, had grounded knowledge and meaning in biological evolution. Dennett had sketched an evolutionary account of meaning in *Content and Consciousness,* although philosophers then were more interested in other aspects of his work (see Section iv.a). By the mid 1980s, however, two influential examples of teleological or evolutionary semantics had appeared.

The philosopher of science David Papineau (1947–) argued that the content of beliefs depends on how they guide actions to satisfy desires, whose content is basically determined by natural selection (Papineau 1984, 1987). Similarly, Millikan (1933–) grounded intentionality in evolutionary history (Millikan 1984). Her book title was deliberately provocative: *Language, Thought, and Other Biological [sic] Categories.* This was guaranteed to raise philosophical hackles in devotees of the later Wittgenstein, and neo-Kantians in general (see Sections vi-viii, above).

Millikan upset many science-inclined naturalists too, by giving more philosophical weight to evolution than to neuroscience. Thus she argued that a perfect simulacrum of a human being, magically constituted in the middle of a swamp by a sudden combination of the relevant molecules, would have *no* beliefs, desires, or other intentional properties (1984: 93, 337f; 1996; cf. Boorse 1976). It would, of course, utter the very same words as a human being would, if engaged in 'conversation.' For all the language-relevant events in the swamp-man's brain (and ears, and lips ...) are, by hypothesis, identical with those of a person. But it wouldn't be a genuine conversation–for, on the swamp-man's side, no meanings or intentions would be being expressed. (In her defence, one could point out that we accept thermodynamics *even though* it allows the theoretical possibility of a snowball in Hell: is swamp-man any more implausible?)

This imaginary example highlighted her central–and controversial–claim, that current meanings depend in part on events that happened millions of years ago. Millikan was saying, in effect, that Searle had been wrong about the "something more" that's needed *in principle* for intentionality. According to her, it's not neurochemistry as such that grounds meaning–nor even neurochemistry in interaction with the body and environment. Only evolutionary history can fix the system's semantics.

If Millikan's (or Papineau's) version of biological semantics is correct, then no 'ready-made' AI-system, nor even a self-organizing–but non-evolutionary–A-Life system, could enjoy mind, intelligence, or meaning.

However, evolutionary semantics was later related to research in evolutionary robotics (Boden 2001). We saw in Chapter 15.vi.c that a robot's neural-network 'brain' may evolve 'feature detectors' analogous to those found in mammalian visual cortex. So a mini-network may evolve that's sensitive to a light-dark gradient at an orientation matching one side of a white cardboard triangle, and that's used by the robots as a navigation aid (Harvey, Husbands and Cliff 1994; Husbands, Harvey and Cliff 1995). Such examples challenge Searle's (1980) view that the "meaning" of a computer model must always be derivative, and arbitrary to boot (see Section v.c).

One could debate whether the feature detector means "light-dark gradient sloping up and to the right" as opposed to "left side of the white triangle". But similar difficulties attend the ascription of non-conceptual content to animals. (Are bug detectors really *bug* detectors, whether for the frog or for the frog's brain?–see Chapter 12.x.f and Cussins 1990: 416f.).

The important point is that the various meanings one might want to ascribe to the robot aren't arbitrary. Nor are they derivative, based only in the human purposes involved in their design. They aren't based purely on causal regularities, either. They spring to mind as candidate meanings because the mini-networks concerned have evolved, within that task environment, to discriminate certain visual features and guide the robot's movements accordingly. That is, they're environmentally, enactively, and evolutionarily grounded.

However, to say these A-Life "meanings" aren't arbitrary isn't to say they're genuine. There's no consensus among A-Life researchers on whether evolutionary robotics could produce real intentionality. For the pure A-Life functionalist, it could: the triangle detector is a primitive case, and more advanced (animal-like) examples would embody richer meanings. For Maturana, it couldn't: evolution and intentionality can occur only in biological organisms–so quasi-evolved robots can quasi-embody only quasi-meanings.

Nor is there a consensus among philosophers unconnected with A-Life, for the nature of life, mind, and the life-mind relation remain controversial.

Not everyone accepts an evolutionary semantics, for example. A causal semantics can't support the commonsense intuition that mind can arise *only* from life, unless the relevant causal relations can be shown to arise *only* in living things. And a model-theoretic semantics can't support it at all.

The competing A-Life methodologies of the early 1990s were systematically compared, and related to earlier philosophies of life, by Peter Godfrey-Smith (1994). He distinguished three dimensions of variation: internalism and externalism; asymmetrical and symmetrical externalism; and weak and strong versions of the continuity of life and mind.

Internalist approaches see life as autonomous self-organization, wherein internal constraints govern the history and interactions of the constituent units of the system. Examples include autopoietic theory and Stuart Kauffman's autocatalytic networks (15.viii.b). Externalist approaches explain the system's internal structure primarily as a result of its adaptive interactions with the environment. Work on evo-

lutionary robotics is one example.

The asymmetric externalist emphasizes the organism's adaptive responses to its environment. By contrast, the symmetric externalist pays attention also to the active role of the adaptive organism in shaping that environment. Examples are situated robotics, and Ray's or Pattee's models of co-evolution, respectively.

Finally, the weak continuity theorist sees mind as emerging only from life, but as significantly different from it, whereas the strong continuity theorist regards mind and life as ontologically similar, sharing basic organizational principles. Descartes wasn't a continuity theorist at all, for he saw mind and living bodies as utterly distinct (see Chapter 2.iii and Matthews 1977). Examples of strong continuity theorists include the *Naturphilosophien* (Chapter 2.vi), the cybernetics movement (Chapter 4.v-vii), and autopoietic theorists. Arguably, they also include philosophers of non-conceptual content (Chapter 12.x.f) and participatory computation (see Section ix.e, above). And someone who argues that not all living things are cognitive systems (see above), is supporting weak continuity in that respect.

However, "mind" covers a number of abilities, and some of these may be strongly continuous with life whereas others aren't. Language has often been seen as a cut-off point. For instance, we saw in Chapter 2.ii.a,g that Aristotle was a strong continuity theorist for perception and autonomous movement, but perhaps not for human reason (cf. Matthews 1992). Heideggerians who confine *dasein* to human beings, or Wittgensteinians who ascribe intentionality only to linguistic concepts, count thus far as weak continuity theorists. But some neo-phenomenologists (such as Clark and Wheeler) ascribe intentionality to non-human animals, too.

Analogously, many AI connectionists allow that GOFAI insights will be needed to model the 'logical' aspects of human thinking (see Chapter 12.viii-ix), whereas some dynamical theorists deny this (Section vii.c, above). And *nouvel AI* (a label recalling the minimalism of *nouvel cuisine)* insists that AI must be grounded in 'lower' abilities, like those of our evolutionary precursors, whether or not it has to add GOFAI methods on top.

In sum, the relation between life and mind is still highly problematic. That applies to work in AI/A-Life, and to philosophy too. The commonsense view is that the one *(life)* is a precondition of the other *(mind).* But there's no generally accepted way of proving that to be so.