Proceedings of the International Symposium on AI-Inspired Biology, Jackie Chappell, Susannah Thorpe, Nick Hawes and Aaron Sloman (Eds.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

CHAPTER 17: WHAT NEXT?[1]

Margaret A. Boden [2]

Abstract.

This is a re-formatted version of Chapter 17 of M. A. Boden, *Mind As Machine: A history of Cognitive Science (Vols 1–2),* Oxford University Press, Oxford, 2006. The chapter title is *WHAT NEXT?* which is also the topic of the final discussion session of the AIIB symposium, which Professor Boden has agreed to introduce.

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

It is included on the AIIB web site with the kind permission of the author.

1 Introduction

This chapter might have joined Chapter XI of *Through the Looking Glass* as one of the shortest in the English language. For in response to "What next?", what is there to say but "Who knows!"? Fundamental advances, in particular, are unforeseeable. As Captain Cook's biographer put it, "Genius, of whatever sort, takes us unawares: is not, even in retrospect, deducible" (Beaglehole 1974: 3).

One doesn't have to adopt a Romanticist view of creativity, nor a literal interpretation of "genius", to agree with that. Creative ideas are unpredictable for a number of very different reasons, not all of which will be mentioned here (but see Boden 1990a: chap. 9). And some are more unpredictable than others. Even a carefully designed technological artefact will have some unpredictable features, as we saw in Chapter 8.v.b (Ziman 2000b). A relatively speculative creative idea can be more surprising still. (Its social impact is even less predictable: Tim Berners-Lee himself couldn't have foreseen that in a mere three years, from 1993 to 1996, the Web would grow from 130 sites to over 600,000–Battelle 2005: 40.)

More precisely, *historically* creative ideas, never generated by anyone before, are unforeseeable. Creative ideas that are new only for the person concerned can sometimes be foreseen, and even deliberately brought about, by other people–think of Socratic dialogue, for example. In what follows, I'll use "creative" to mean historically creative.

17.i: What's Unpredictable?'

One source of unpredictability is serendipity: the finding of something valuable without its being specifically sought. Since this is unexpected by definition, prediction simply isn't on the cards when it's involved. The classic case in science is Alexander Fleming's noticing the dirty dish of agar-jelly, which led eventually to the discovery of penicillin. In cognitive science, the part-accidental discovery of several visual feature detectors is another illustration (Chapter 14.iv.a-b). In both cases, of course, a good deal of careful and systematic work had to follow the initial observation, before anything worth calling a "discovery" could be achieved.

Another source of unpredictability is change in the wider cultural context. The generation–and still more, the acceptance–of new scientific ideas can be discouraged, encouraged, and even part-guided by social/political factors (see 1.iii.b-d and 2.ii.b-c). The positive reception of heterarchy (10.iv.a) and distributed cognition (13.iii.d-e), for instance, was influenced by political ideology, and of expert systems by nationalism and economics (11.v). Even if cultural changes could be predicted (which they can't), their effects on contemporary scientific thinking could not.

This applies also to shifts in the power relations between the various groups/disciplines within cognitive science, and so to *what will be seen as cognitive science* in the future:

> Some 25 years after its various beginnings there still is no such thing as a core cognitive science. Depending on where one looks, which departments one queries, who one's friends are, the core of cognitive science will be asserted to be neurophysiology, psychology, artificial intelligence, linguistics, or some more vague concept like human/machine interaction or symbolic or connectionist modeling. The result may not have been cognitive science, but it has been exciting and scientifically fruitful. It has created a community of interests and increased interdisciplinary communication. But as of now there are still viable independent cognitive sciences such as neurophysiology, linguistics, and psychology that flourish with or without the cognitive science label or affiliation. *It is difficult to say at this point where this will lead* (Mandler 1996: 23; italics added).

A specially important type of cultural change concerns new technology. Sometimes, the new instruments are needed in order to do things that obviously needed doing: for instance, developing microelectrodes to record from the cell-body of a single neurone–or anyway, a very small number thereof (2.viii.e). Such cases are relatively predictable: it was clear that people would try, and probable that someone would eventually succeed. Other technological tools may come as more of a surprise, at least to people in other areas of science: a few biophysicists may have been able to predict brain-scanning techniques, but psychologists couldn't.

Technology can be used to prompt new concepts, as well as to find new data. Indeed, technical ideas have been transmuted into psychological theories on many occasions (Gigerenzer 1991b, 1994). Moreover, machines have been used as analogies for the brain for hundreds

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

of years (Fryer 1978)–juke-boxes included (2.viii.f). The latest, of course, is the computer itself. So sceptics often say that it's just the latest in a long line of such analogies, to be displaced eventually by some unforeseen invention coming who-knows-when.

Well, yes and no. In cognitive science, the computer isn't merely a superficial analogy, a metaphor fished out of the memory–perhaps for purposes of popularization–after the real scientific work has been done. On the contrary, it provides substantive concepts in psychological and neuroscientific theories.

The computational concepts concerned were diverse even at the outset (4.ix), and have multiplied over the years (16.ix). Besides symbolic, connectionist, and evolutionary AI, they include dynamical systems described by differential equations (14.vi, 15.viii-ix and xi). The future may well hold unpredictable new machines, even less imaginable today than quantum computers are. But cognitive scientists believe that only *some sort of computational machine* will be relevant. For their key claim is that mind can be explained (not by today's ideas about computation, but) by *whatever theory turns out to be the best account of what computers do* (Chrisley 1999; see 16.ix.f). In that sense, they would endorse Philip Johnson-Laird's remark that: "The computer is the last metaphor; it need never be supplanted " (1983: 10).

Other psychological reasons for unpredictability apply to all instances of creativity. They include the rich idiosyncracy of human minds and the relative–though only *relative* –freedom of creative thinking (13.iv; cf. 7.i.g). Introspectively, it may seem as though almost anything can happen–at least, according to the molecular biologist Francois Jacob:

> Day science employs reasoning that meshes like gears .... One admires its majestic arrangement as that of a da Vinci painting or a Bach fugue. One walks about it as in a French formal garden ... Night science, on the other hand, wanders blindly. It hesitates, stumbles, falls back, sweats, wakes with a start. Doubting everything ... It is a workshop of the possible ... where thought proceeds along sensuous paths, tortuous streets, most often blind alleys (Jacob 1988: 296).

If the creative scientist himself "wanders blindly" much of the time, so much less can his thoughts be foreseen by other individuals–who don't even know his *present* thinking in much detail.

Moreover, some new ideas strike us as paradoxical, not to say crazy, even *after* they've occurred. That often happens in instances of "transformational" creativity, in which one or more dimensions of the previously-accepted style of thinking is/are radically altered or dropped (Boden 1990a: chaps. 3-4). The more basic the dimension, the more fundamental the conceptual change will be. In such cases, it's hard for the new idea to be understood, and even harder than usual for it to gain acceptance. *A fortiori* it's harder to predict.

In particle physics, that's par for the course. Freeman Dyson reported an encounter between Niels Bohr and Wolfgang Pauli, who'd given a lecture on his new theory:

> Bohr rose to speak. "We are all agreed", he said to Pauli, "that your theory is crazy. The question which divides us is whether it is crazy enough. My own feeling is that it is not crazy enough" (Dyson 1958: 74).

And Dyson commented:

> The objection that they are not crazy enough applies to all the attempts which have so far been launched at a radically new theory of elementary particles. It applies equally to crackpots. Most of the crackpot papers which are submitted to *The Physical Review* are rejected, not because it is impossible to understand them, but because it is possible. Those which are impossible to understand are usually published. When the great innovation appears, it will almost certainly be in a muddled, incomplete and confusing form. To the discoverer himself it will only be half-understood; for everybody else it will be a mystery. *For any speculation that does not at first glance look crazy, there is no hope* (italics added).

Cognitive science as a whole is less *rococo,* less conceptually bizarre, than particle physics. But several seemingly crazy ideas have found a respected place within it, after being fiercely resisted as "absurd"–perceptual expectancies, for example (Chapters 6.ii and 16.v.f), and object-oriented programming (10.v.d and 13.v.d). And remember the punchline of the quip about *Plans and the Structure of Behavior:* " ... and Pribram believed it!" (6.iv.c). When people said that, they weren't dismissing those new ideas as worthless. They were allowing that they were weird-but-interesting, so worth thinking about. (Karl Pribram was made the fall guy because he'd recently defended a holographic theory of memory–hardly the usual bread-and-butter fare: 12.v.c. Nor is Bergsonian philosophy, but Pribram later dallied with that as well: 16.x.a. His reputation as a maverick was deserved. However, even those who called him "crazy Karl Pribram" later admitted that his strange ideas had "caught on," and that "his neurophysiological speculations are decades beyond other physiological work"–Walter Weimer, interview in Baars 1986: 309f.)

Many future contributions, too, will seem weird initially–though just how weird they'll need to be remains to be seen. The "particle physics" of the field is conscious experience. This has already prompted many highly counterintuitive theories, including some crackpot publications. I argued in Chapter 14.xi.e that a currently undreamt-of (i.e. crazy) approach will be needed to explain it.

Close runners-up in order of difficulty, and so in licensed craziness, are intentionality and computation. We saw in Chapter 16.ix.e how an extraordinary (crazy?) theory of those-two-together has come from an AI scientist/philosopher who thinks that "For sheer ambition, physics does not hold a candle to computer or cognitive ... science" (Smith 2002: 53).

Sometimes, experts declare future progress to be not so much unpredictable as impossible. This view was implicit in Thomas Watson's notorious remark in 1943, as IBM chairman, that "I think there is a world market for maybe five computers." (He died in 1956, so never knew just how wrong he was. But he wasn't alone: Howard Aiken, of all people, said "there will never be enough problems, enough work, for more than one or two of these computers"–Edwards 1996: 66.) And it was explicit in the advice given to Konrad Zuse in 1937 by Kurt Pannke, a manufacturer of specialized calculators:

> "Someone informed me", Dr. Pannke began, "that you have invented a computing machine. Now, I don't want to discourage you from continuing to work as an inventor and from developing new ideas, but I must go ahead and tell you one thing: in the field of computing machines, practically everything has been researched and perfected to the last detail. There's hardly anything left to invent ...." (Zuse 1993: 42).

(To be fair to Pannke, he later changed his mind. He provided money to fund Zuse's home-based research, and recommended his machine to the German military–fortunately, with no effect: see 11.i.a.)

17.ii: What's Predictable?

I can't imagine anyone suggesting that there's "hardly anything left" to be discovered in cognitive science. But I've just allowed that creative ideas can't be predicted, only awaited. So perhaps I should now present you with an empty page, and leave it at that? After all, that's a respected rhetorical device. Laurence Sterne did it 250 years ago, when he declined to describe a beautiful woman in *The Life and Opinions of Tristram Shandy,* leaving it to the reader's imagination instead.

I don't have the courage to follow Sterne's example. But it wouldn't be appropriate in any case. For there is something that can be said.

All scientific research, in whatever domain, is located within some identifiable conceptual space where further creative exploration (and transformation) is clearly possible, and where some dimensions seem especially rich in potential with respect to current unsolved problems (Boden 1990a, 2004). Peer-review, especially of proposals for future research, depends on that fact. We can't predict the detailed outcome of such explorations and transformations, much as David Livingstone couldn't foresee his discovery of the Victoria Falls. But we can reasonably expect that if we follow *these* dimensions of the space (compare: the Zambesi river, the mountains glimpsed ahead ...), we'll find something of interest. That is, we can have intellectually defensible, if not infallible, hunches about where future discoveries are likely to occur.

Insofar as such predictions are possible, I've indicated mine already. The previous chapters have told "the story so far"–but always with an eye to possible future episodes. So the relatively small volume of *recent* work that I've mentioned was chosen not just because it's recent, nor even because it's intriguing. It was selected because I think it's promising, capable of development in ways that seem likely to be fruitful.

One way of justifying our hunches about where interesting new ideas are likely to arise is to rely on sub-hunches about *how* those ideas might be generated. In other words, some specific exploratory pathways are recognizable as familiar ones, because they've often been found to be fruitful.

- For instance, once a simple deterministic space has been defined, it's very likely that people will eventually try to complexify it in certain ways. So when John von Neumann defined the basic cellular automaton, he knew very well that probabilistic and even evolutionary CAs would be explored later (Chapter 15.v-vi).
- Again, once problem solving had been seen as a simple hierarchy (6.iii), it was inevitable that more complex and/or 'open-execution' plan hierarchies would be explored (10.iii.c). It was even a good bet that theories of problem solving would eventually be transformed by hierarchy's being made less pure (10.iv.a), or perhaps deliberately dropped (13.iii.b, 15.viii.a).
- Third, when Alan Turing wrote his morphology paper, he knew that increasingly complex systems of diffusion-reaction equations would be explored, once computer power allowed (15.iv).

(Similar remarks apply to creativity in artistic contexts. So, for example, it was nigh-inevitable that post-Renaissance composers would progressively complexify tonal harmony. And it was always on the cards that someone–it happened to be Arnold Schoenberg–would eventually transform the space of tonal music by dropping the home-key constraint altogether: Rosen 1976; Boden 1990a: chap. 4.)

In short, the common notion that creative thought is unpredictable because it's chaotic (in the everyday sense) is mistaken. There's significant method in creative madness. It's our tacit recognition of this fact which enables us to identify certain work as promising, even though we can't spell out the promises.

17.iii: What's Promising?

The recent empirical research that I see as promising in these terms includes the following (listed here in no particular order):

- hybrid systems I: symbolic/connectionist (Chapters 12.iii.d and ix.b, 15.viii.a)
- hybrid systems II: situated/deliberative (7.iv.b and 13.iii.c)
- hierarchical networks (12.viii.b and ix.b)
- connectionist work on the role of imagery-of-words (12.ix.e)
- statistical approaches to NLP (9.x.preamble and 9.x.f)
- integration of connectionist learning with detailed neurophysiological data (14.ii.d)
- modular and/or time-based neural networks (12.ix.a, 14.ix.g)
- programmed/evolutionary neuromodulation (14.ix.f)
- AI-evolved organic-silicon computing networks (14.ix.f)
- computational neuroethology (14.vii, 15.vii)
- insect navigation strategies (15.viii.a)
- types of cerebral representation, especially emulators (14.viii)
- the epigenesis of thought and language in normal and brain-damaged children (7.vi.g-i)
- models of clinical apraxia and aphasia (12.ix.b, 14.x.b)
- theories of control in hypnosis (7.i.h)
- brain-scanning, *provided that* it's related to specific psychological theories (14.x.b)
- developmental trajectories (12.viii.c-e and x.e)
- fast/simple heuristics (7.iv.f-g)
- the origin of specific bounds on human rationality (7.iv.h)
- cognitive technology, including virtual reality (10.i.h, 13.vi, 16.vii.d)
- computational theories of creativity (9.iv.f, 13.iv)
- evolutionary modelling (14.ix.d and f, 15.vi)
- achieving open-ended evolution and/or creativity (13.iv.c, 15.vi.d)
- mathematical analyses of dynamical systems (14.vi and ix.b, 15.ix.b and xi.b)
- homeostasis in CTRNs (15.xi.a)
- distributed cognition and agents (8.iii, 12.ii-vi and x, 13.iii.c, 15.viii-ix)
- computational architectures integrating knowledge, motivation, and emotion (7.i.e-g and 7.iv.b-c).

If forced to choose only one of these items, I'd pick the last: work on integrated mental architectures. Indeed, I did that on the 50th anniversary of the 1953 discovery of the double helix, when the British Association invited several people to write 200 words for their magazine *Science and Public Affairs* on "what discovery/advance/development in their field they think we'll be celebrating in 50 years' time". This choice reflected my own

longstanding interests in personality and psychopathology (Preface.ii). But it wasn't idiosyncratic: two years later, the UK's computing community voted for "The Architecture of Brain and Mind" as one of the seven "Grand Challenges" for the future (http://www.uk.crc.org.uk/Grand_Challenges/index.cfm). One member of the five-man committee carrying this project forward is Aaron Sloman, who's being thinking about architectural issues since the 1970s (7.i.f, 10.iv.b, and 16.ix.c). If progress is to be made on this front, my hunch is that his team will be in a good position to make it.

The Grand Challenges grew out of the UK government's "Foresight" Programme (instituted in 2003 for a ten-to-twenty year planning horizon), and in particular out of its "Cognitive Systems" Project. Naturally, government ministers aren't falling over themselves to help solve the problems of cognitive science for their own sake. For them, applications are all–whether in health, education, business, transport, arts and entertainment, or (of course) the military. But as the Project's official Report (n.a. 2004) makes clear, scientific and technological motives are often very closely related (and can be satisfied only by interdisciplinary thinking). It should be no surprise, then, that architectures to support "emotional" robots and "social" human-computer interactions are now being investigated at the behest of Whitehall–and, naturally, of the Pentagon too.

The strength and range of the list of "promises" given above show that cognitive science is still a fruitful "scientific research programme" (Lakatos 1970). *Mind-as-machine,* in both its incarnations (1.ii.a), has generated many suggestive theories. These have been amended–and sometimes dropped–on the basis of further advances in our understanding, but in many cases the central insights remain. Marr's work on vision is one obvious example (7.v.b-f), but others have been described in previous chapters.

The field's potential won't be unlocked without new psychological-computational *theories.* Greater computer power may well be necessary, but it won't be sufficient. Even quantum computers and hypercomputers won't suffice to fill the bill (16.ix.a). Fundamental scientific advance will need more Ideas, not just more Bytes. Likewise, more and/or fancier PET/fMRI brain scanning won't suffice either, even though it will often be useful (14.ii.d and ix.c).

On a higher plane of abstraction, I've discussed some recent philosophical research concerning:

- the nature of computation (16.ix)
- the variety of virtual machines (16.ix.a)
- conscious experience (14.x-xi)
- the nature of intentionality (16.x.d)
- the origin of conceptual content (12.ix.e and x.f)
- the nature/existence of non-conceptual content (12.x.f and 16.viii.b)
- mind and/as embodiment (16.vii)
- the boundary between self and world (16.vii.d)
- the nature of life, and its relation to mind (15.i and 16.x)
- the resolution/reintegration of neo-Kantian and analytic philosophical viewpoints (16.vii.b-d, ix.d-f, and x.a).

All of these matters will be key foci of effort and controversy in the foreseeable future. Indeed, they're so difficult, and so deep, that I expect them to remain key foci well over a hundred years from now. For as remarked at the outset of Chapter 16, philosophical problems don't get solved in a hurry.

Nor, in these cases, will they get solved in disciplinary isolation. They'll require fundamental *and reciprocal* advances in up to five fields: philosophy, psychology, anthropology, neuroscience, and AI. (Theoretical linguistics, as opposed to the philosophy of language, is less relevant here–unless we include *cognitive* linguistics: see 7.ii.preamble, 9.ix.g.)

17.iv: What About Those Manifesto Promises?

In Chapter 6.iv.c I said that a good way of judging how far cognitive science has succeeded is to compare it with the hopes/promises expressed in *Plans and the Structure of Behavior* (Miller, Galanter and Pribram 1960). By the turn of the millennium, virtually all of MGP's promises had been at least partially met. The "satellite images", and the Newell Test, outlined in Chapter 7.vii surveyed many different examples.

To mention just two:

- hypnosis has been demystified (along with multiple personality and religious experience): (7.i.h-i, 8.vi.b, and 14.x.c), and
- MGP's distinction between Plans as animal instincts and as human purposes is now far better understood (7.i and iv, 14,vi.c, and 15.vii).

Although discussions of these matters have been hugely complicated since they wrote their manifesto, today's answers are broadly consistent with theirs. For TOTE units were–deliberately–defined so abstractly that they covered *both* the inbuilt sensori-motor skills of crickets and hoverflies *and* the deliberative (and hypnotic) planning of human beings.

Neither "demystified" nor "far better understood" implies that all the relevant questions have been answered. Far from it. But we're much clearer now about just how MGP's questions can be profitably put.

Consider, for example, their nature-nurture distinction mentioned above. This simplistic duality has given way–within cognitive science, if not yet in the minds of the general public–to an epigenetic view of development. This view was already waiting in the wings before cognitive science got started (5.ii.c). Now, it's prominent in disciplines as varied as psychology, neuroscience, philosophy, A-Life, and robotics (7.vi, 14.vii and ix.c, and 15.viii.a).

There's no reason why this process should cease now. And it doesn't require every psychological question to be answerable by a simulation. For MGP's futuristic remarks concerned a general approach to the mind: computational theorizing, not necessarily computer simulation as such. We'll surely see many new computer models (some of which will reflect new findings in neuroscience). We'll probably see radically new *types* of model (16.ix). Functioning computer models can test a theory's implications and coherence more rigorously than any other method (7.iii.c). But the novel theoretical concepts are what's important, in understanding what sort of system, or machine, the mind/brain is.

One thing is beyond dispute: that the rich subtlety of human minds is even more awe-inspiring than the arch-humanist Wilhelm von Humboldt (9.iv) believed it to be. Indeed, I've already identified this realization as the major result of computational psychology as a whole (7.vii.a). It follows that it will never be possible to capture

every psychological detail, whether in a theory or a simulation. Predicting, explaining, or interpreting the specific thoughts/actions of individual people will always be largely "idiographic" (7.iii.preamble), a matter for the unargued intuitions of psychologists *qua* human beings, not for their deliberate conclusions *qua* scientists.

However, that doesn't spell disappointment for MGP. For on the one hand, idiographic insights can often be enriched, and sharpened, by considering general mechanisms. Remember, for instance, the varied ways of expressing different types of anxiety in speech (7.ii.c). On the other hand, the prediction/explanation of highly particular personal matters wasn't what MGP were aiming for. (Nor is this the aim of scientific psychology in general: 7.iii.d.) Rather, they hoped to understand how such phenomena are *possible.*

It's not only MGP's questions which can now be posed more fruitfully. The familiar puzzles that opened this story (1.i.a), many of them centuries old, have all been illuminated–and some even solved–by the successors of the visionary manifesto.

More answers will doubtless be found: the future of cognitive science will be as exciting as its past. But to say *what they'll be* would be like an eighteenth-century Admiralty Board foreseeing James Cook's extraordinary achievements in navigation and map-making: impossible.