# On the creation of a mind

*Andrew Brown, Peter Wilson, Mark Zwolinski and Bashir Al-Hashimi*
*Department of Electronics and Computer Science*
*University of Southampton*

## Preamble

It could well and easily be argued that this proposal has no place here; and that it could and should be part of the UKCRC Grand Challenge "The architecture of Brain and Mind", The synopsis of the 2002 meeting is attached - almost all of it is informative and relevant and addresses many of the direct questions in Steve Furbers' call for proposals, which is why we do not intend to reproduce it. We're all busy people and busy people hate documents longer than six sides. Read the first page of Mike Denhams' synopsis. Alternatively, type "architecture brain mind" into your favourite search engine and stand back.

## Why here/now/this?

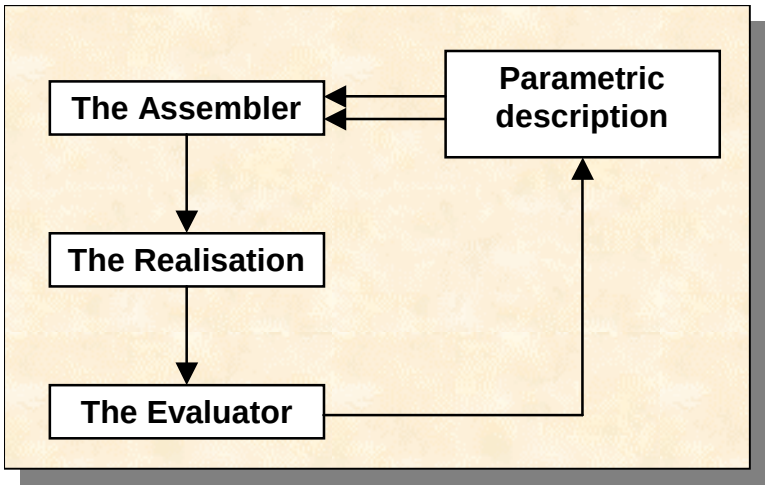
- Prosaically, we're not on the UKCRC mailing list; so why is that stopping you, just blag your way in, it certainly wouldn't be the first time, we hear you cry. Unquestionably true, and we probably will.

- This notwithstanding, a project of this breadth and depth can never have too many communities engaged, and we perceive that the microelectronic design community is under - if at all - represented in that particular challenge. It is possible we may offer a perspective that is different.

Maybe what follows contains too much detail. Maybe it's all been thought of before. Sorry.

## The big picture

Cutting to the gunfight, we'd like to build a mind. Not necessarily a human or even mammalian mind, largely because we don't know how to distinguish one type from another, or, more to the point, even recognise one when we see one. (Anyone who has tried to teach, for example, Pt II Numerical Methods to electrical engineers will have experience of this.) The system we would like to build is outlined in the figure below.....



It's a feedback system, and from a wide enough perspective, feedback systems can train things to do things. But we don't want to build a neural net and train it to say "Thanks" when we offer it a biscuit. We want to offer it the 383rd biscuit and hear it say "No thanks, I'd rather play chess". Emergent behaviour....

## The Assembler

***Self organisation***
There is a view, from an information theoretic perspective, that holds that when sperm meets egg there is simply insufficient information present to describe the outcome. The information is encoded and compressed, which accounts for the variability of the offspring. A degree of self-organisation is evident. And yet - engineering undergraduates notwithstanding - humans do not give birth to monkeys; dogs don't have kittens. It's not *rare*, it *never* happens. And yet the physical similarities between the organisms are overwhelming. The parameter space describing viable organisms has deep local minima, which are all very close 'sideways', yet with impossibly high 'walls' between.

***Another point of view***
Self organisation on this scale and at these levels of complexity just does not hold water. Diverse and imprecise though cell-based life on this planet is, it is the *only* complex system that (supposedly) organises itself like this. Piles of sand, ant colonies, cellular automata and flocks of birds are, above a certain level of granularity, homogeneous. Living organisms are not. A school of thought is emerging that holds that the DNA of eukaryotes (organisms consisting of cells with a nucleus) holds much more information than previously thought.

- In eukaryotes, ~95% of the DNA is considered "junk" - it does not code for any of the known 21 proteins and the accepted wisdom is that the junk is pruned out as a precursor to protein synthesis, smashed up and resorbed. It is there probably as the disabled legacy of failed mutations or viral attacks.

- The junk appears to be far more stable across multiple generations than would be expected if it were genuinely information free.

- Externally, most high level organisms are bilaterally symmetric, but internally they are not. With very few exceptions, humans have their heart on the left, appendix on the right. If self-assembly dictates geometry, why is the population not 50:50?

- With very little modification, virtually every creature that has ever existed would be viable today. But after the species explosion of the Cambrian era about 525 million years ago, the *rate* of creation of new species fell dramatically.

- The species explosion corresponded with the appearance of eukaryotes. Prokaryotes (the dominant, but extremely simple life form before then) employed (and still do employ) a different metabolic pathway to extract information from their DNA: introns (junk) carry a significant metabolic cost

in prokaryotes and so short term evolutionary pressures select against it. The overhead in eukaryotes is negligible.

- To build any complex mechanical or electrical system, you need to specify two things: the *components* used to build the system, and the *instructions* dictating how those components are to be assembled. In general, the size of the instruction set grows polynomially with the size of the system. Most biological systems are orders of magnitude more complex than any man made creation; to assume or expect self-organisation could be construed as a giant leap of faith?

- In summary, the idea here is that the protein coding DNA (exons) describe the components - a relatively mature view - and the junk DNA (introns) describe the assembly instructions for the diversified cells.

- Why cannot this theory - or a variation of it - apply to the neural organisation in the brain?

***Incrementalism***
The UKCRC papers make much - quite rightly - of the enormous advances in understanding the function of neural systems at various levels of granularity, and also the growth of computational power available. We're certainly not detracting from this massive body of work, but we do wonder if there is another way into this particular problem.

***A gedanken experiment***
Imagine you know an awful lot - all there is to know - about electronics at the analogue component level, but have never heard of discrete logic and know nothing about digital computers in any form. You are given a flat (non-hierarchical) transistor level circuit diagram of a twin processor Pentium V PC - yes, including the disc drive and screen, keyboard and so on. You've also got a working example, which is playing DOOM, or the mindlessly violent game of your choice. You're allowed to press the buttons and wobble the mouse, and you can take the back off and prod at the boards with a 30 year old AVO mk 8, but nothing more sophisticated.

How long will it take you to deduce the existence of
a) digits
b) the von Neumann fetch-execute cycle
c) the concept of software

***Two supplementary questions:***
How many neurones can you kill in a mind before anyone notices?
How many wires can you cut in a PC before anyone notices?

***Throwing the dice:***
For years, people wondered how and why birds flock. The flocks are generally stable - they don't fall apart, and the individual birds don't collide. Yet there is (almost) no non-nearest neighbour interaction, and no Ubermind. One day, someone sat down with a 3D cellular automata program, and started to play. Now no self-respecting sci-

fi film cityscape is complete without flocks of computer generated birds wheeling around the skyscrapers.

Take a (large) number of NAND gates. Connect them up randomly, and vary the statistics (fan in/out) of the connectivity. Most of the time, boring, inert collections of gates are formed. But the parameter space contains 'islands', where finite state automata - of varying degrees of complexity, some quite big - form. And what is a mind, after all, if not a (large) finite state machine?

Take a slab of brain. Large enough to maintain some functionality, small enough to be reasonably homogenous. *Unsurprisingly*, behaviour implies a certain statistical connectivity. Build an individual neurone model and a simulator and a random assembler. *Surprisingly*, statistical connectivity implies behaviour. Maybe the statistics give us an initial condition and we iterate to sentience?

Maybe it really is that simple?

## The Realisation

We get a little less Zen here: Ultimately, only purpose-designed hardware can deliver the massive parallelism essential for delivering the necessary throughput. However, being realistic, conventional software simulation is the only way to start. Simulators can be instrumented to the $n^{th}$ degree, and stopped, started and replayed at will. An Itanium PC has a memory space of $2^{64} \approx 10^{19}$ words. Representing the state of a neuron with 100 words = 6400 bits gives us a playpen of around $10^{17}$ neurone instances, ignoring codespace. The numbers are OK.

## The Evaluator

I am on very thin ice here:

- If you take a new born "mind" and deprive it of all stimulus, it breaks.
- Which means there exists a feedback loop somewhere that requires stimulus and supports learning.
- Which means there is a "figure of merit" aka penalty function to drive the feedback. The figure of merit describes a "happiness surface" on which we hunt maxima.
- Is the happiness surface built in? If so, where is it? (But back to the recognising the transistor level von Neumann problem.).
- Evaluation is education.
- But if you start at t=0 (and I can't believe people haven't tried this) why can't you get the monkey to play Rachmanninov?

***A momentary lapse to philosophy - or possibly ethics:***

One incarnation of "simulation" of a bit of software can be the execution of the software itself.

- What if it all works?
- What if we simulate a mind that passes all the evaluation criteria?
- Have we *created* a mind?
- Can we then - legally, morally, ethically - switch it off?