

GRAND CHALLENGE PROPOSAL

Architecture for a mind: requirements and designs

Aaron Sloman

<http://www.cs.bham.ac.uk/~axs/>

The challenge:

Evolution has produced myriad organisms that process information of many kinds in many ways.

These include:

- (a) controlled production of individual organisms from eggs or seeds,
- (b) evolutionary processes storing useful information for future organisms,
- (c) emergent interactions among many individual organisms such as termites building their cathedrals, and
- (d) most recently, large, physically integrated, “expensive”, multi-functional minds and brains in individual organisms, e.g. in birds, apes and humans.

Our challenge is to understand and replicate type (d): *What are the requirements for building powerful, integrated, minds and what sorts of designs can meet those requirements?*

Background:

The idea of building something like a person has fascinated people for centuries: e.g. the “golem” idea, many mechanical toys, Frankenstein, films about man-made monsters, and most recently AI-based synthetic agents in computer entertainments. This requires merging two ancient strands of engineering:

- (1) mechanical control mechanisms (e.g. speed governors, mechanical looms, music boxes, lathes, card sorters),
- (2) machines operating on abstract entities (e.g. calculators of various kinds, machines that operated on both numerical and non-numerical census information and eventually also machines that operate on their own programs).[1]

Electronic technology brought these two strands together in an ever-accelerating process of development both of physical devices and virtual machines to run on them, making it possible to start taking seriously the possibility of replicating human mental capabilities of many kinds thereby giving birth to AI in mid century.

Artificial Intelligence was mis-named, for, besides aiming for new artefacts, many AI researchers also hoped to develop a new science of mind, extending disciplines like psychology, linguistics, neuroscience and philosophy by introducing new forms of explanation, using new ontologies (e.g. processes in virtual machines). Unfortunately, absurdly optimistic predictions were made, partly because the protagonists were untrained in disciplines outside engineering and mathematics: e.g. they did not know enough psychology, linguistics and philosophy.

Turing, more modestly and more sensibly, predicted in 1950[2] “*that in about fifty years time it will be possible to programme computers with a storage capacity of about 10^9 to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning.*” I.e. 30 % of the population can be fooled for five minutes! He also described the question ‘Can machines think?’ as “*too meaningless to deserve discussion.*”

Where next?

Having explored many varieties of machines and varieties of tasks since Turing's time we have now learnt enough to replace the unproductive question “Can machines think?” with a pair of new deeper questions on which real progress can be made:

- What types of machines are there (including types of virtual machines)?
- what kinds of “thinking” can different types of machines do or not do?

To address these we need to expand both our exploration of types of information processing architectures, and the variety of types of mental processes found in humans and other animals, namely processes involving the acquisition, manipulation, derivation, storage, communication and use of *information* of many kinds, e.g., factual information, control information (including skills, motivation and emotions), and meta-information about information. In short, we can explore a space of possible designs and a space of possible sets of requirements along with their matches and trade-offs. Philosophy becomes engineering.

The present context — almost total fragmentation:

For both technical and social reasons, the study of natural and artificial intelligence is now badly fragmented. Investigators look at small sub-systems or small sub-problems. They specialise on vision, on emotions, on motor control, on smell, on reinforcement learning, on language, on a particular stage in infant development, on particular brain functions, etc. Or they focus on particular forms of representations and techniques, e.g. logic and theorem proving, neural nets, behaviour-based systems, evolutionary computations, dynamical systems. But most don't try to combine their models or their mechanisms. Vision researchers ignore work on natural language processing and *vice versa*. Robots are given capabilities, but not the ability to understand or discuss those capabilities. Of course, we need many highly focused specialised researchers, but there is a risk of producing *only* systems that cannot be combined in integrated systems.

One consequence may be a failure to design good interfaces for systems that have to interact with humans, because we don't understand the variety of forms of processing humans use and the variety of tasks they use them for.

The proposal:

Given everything that has been learnt from AI, computer science, software and electronic engineering, psychology, neuroscience, ethology, etc. about the various pieces of the puzzle, and given the huge advances in available computing power, electro-mechanical devices and perhaps soon also nano-technology, the time seems right for a fresh attempt at integration, by putting the pieces together in a working robot, perhaps with the capabilities of a somewhat simplified five year old child. There are three main tasks to be pursued in parallel:

1. Specifying *requirements* for the design to meet (which will include both (a) discovering what the requirements are and (b) designing appropriate formalisms to express them.)
2. Producing a *design* for a system to meet the requirements, including both a design for the physical body and a design for the virtual machine (mind).
This may require extending current design formalisms to support the task.
3. *Implementing* examples of such designs, some using physical simulations and some using advanced robot technology.

Different teams should be allowed to adopt different approaches, provided that they meet regularly at workshops, in an atmosphere combining healthy competition and information exchange.

Difficulties – formulating requirements:

Besides the many obvious difficulties, there is a subtle but huge obstacle that often goes

unnoticed (a form of “ontological blindness”^[3]), namely the difficulty of understanding the *requirements* of the task. In particular we underestimate the difficulty of discovering the capabilities of a child (or any other animal).

For instance, people who think of perception as simply providing information about *physical* properties of the environment or about reliable correlations between *image patterns*, often ignore more subtle and abstract perceptual functions, namely perception of causal powers, like “affordances”, which are *relational* features involving the organism’s (or robot’s) needs and capabilities *as well as* structures, possibilities for change, and causal powers in the environment. How an organism perceives *graspability* will depend on typical needs and goals of that species and also their specific grasping capabilities – using teeth, fingers, or tail, for instance. Vision researchers may spend a lifetime studying simpler, more concrete tasks, such as visual pattern recognition and pattern correlation, ignoring ontologies and formalisms needed to express affordances. *A conjecture: until we understand human abilities to grasp and use visual affordances we shall not understand or be able to replicate some of the most powerful forms of thinking, learning, and problem solving employed in many domains.*

Moreover, some researchers underestimate the perceptual needs of social animals (e.g. the ability to perceive intentions, moods, personalities, in others). When researchers ask: ‘*What mental ontology does a child (or chimp) develop, and how is it encoded in visual mechanisms?*’ they often consider an impoverished range of possible answers, and this “ontological blindness” limits the architectures, mechanisms, formalisms, and assessment criteria considered.

Likewise many of those who study affective states tend to focus on shallow observable *expressions* of emotion rather than the deep information-processing functions of affective control states (e.g. in long term grief or jealousy). Researchers studying communication may not notice the differences between communications about readiness to mate or the location of food and communicating an understanding of transfinite set theory or how to debug programs. The latter requires very different architectures in teacher and learner from those that suffice for the former. Even some insects do the former.

Towards a solution – Architecture-driven exploration:

We are nowhere near explaining or replicating most of the capabilities of a young child dressing and undressing dolls, a squirrel attacking a garden peanut dispenser, a nest-building bird, etc., let alone explaining how humans can design jet airliners, prove theorems about transfinite ordinals, discuss philosophical puzzles, and enjoy creating and experiencing poetry and string quartets.

One way to make progress is to take the most promising existing proposed architectures and attempt to find out what they fail to explain. For example there are theories proposing various kinds of interactions between different concurrently active processing layers which differ in:

- (a) the forms of representations used,
- (b) the kinds of semantics and ontologies they deploy,
- (c) the degree or kind of task abstraction,
- (d) the varieties of learning,
- (e) the varieties of control (including motivation and emotions), and
- (f) the extent to which they are inwardly or outwardly directed.

These architectural features map in subtle, complex, and very indirect ways onto external behaviours, making testing difficult. We shall need to develop new ways of deriving properties of the behaviours.

This will probably require new formalisms for describing the complex architectures and the requirements that they have to meet. For instance we shall need new formalisms for representing constraints and possibilities for action (affordances) inherent in a structured 3-D scene at different levels of abstraction corresponding to different sorts of goals.

It would help if we could design a meta-theory for architectures: a way of systematically generating possible architectures covering a wide range of systems. Then when we find that a particular architecture is inadequate to explain some capabilities, e.g. visual problem solving, or enjoyment of games, we can use the meta-theory to generate alternative architectures including different forms of representation, forms of reasoning and forms of control. Producing an adequate meta-theory to support the right variety of architectures is a major challenge.

Within this framework we can hope to develop theories of *possible* trajectories through the space of architectures, and we can try to relate those to *actual* evolutionary trajectories followed across generations, and also individual developmental and learning trajectories. (See [3])

In parallel with such top-down and middle-out explorations we shall need to collaborate both with brain scientists to learn more about information processing mechanisms not yet invented by engineers, with psychologists to find out in more detail what the requirements for a human-like architecture are, and with others to find out about intermediate architectures used by other animals.[4]

As the architectural ideas develop we can present more and more pointed and precise questions to the other disciplines, thereby driving their research. One example is using analysis of possible types of malfunction in a designed architecture to guide empirical research on types of brain damage or disease and their consequences.

As has happened throughout the history of AI, this sort of project will provide demanding new requirements for generic hardware and software computing technologies and will benefit from advances in CS theories and technologies, including advances concerning networked virtual machines, as proposed for the "global computer", described at the sample grand challenges website [5].

Prospects:

Understanding how to build a child-like mind may be a necessary pre-requisite for many other tasks. *Can we ever build a human child-like mind?* Not soon: vast amounts of computing power help little if we don't yet know what the task is, i.e. exactly what we need to model, or to explain.

We also do not yet understand the variety of mechanisms that can be built on all this computing power: Our knowledge of the space of possible virtual machines with complex hybrid architectures combining many forms of representation and mechanisms for operating on them, is still in its infancy. But this project will help to accelerate progress. It will clarify many sub-tasks and provide a launch-pad for decades of further work in several disciplines with potential for applications in robotics, education, therapy, computer entertainments and many intelligent software-systems. Milestones can be set up to direct the work and guide evaluation [6].

Additional ideas can be gleaned from the international Robocup Rescue competitions [7] and from DARPA's cognitive systems project (with which we should collaborate [8]).

Some references:

- [1] A. Sloman, (2002), The irrelevance of Turing machines to AI, in Ed. M. Scheutz, *Computationalism: New Directions*, MIT Press, Cambridge, MA, pp. 87–127, (Available at <http://www.cs.bham.ac.uk/research/cogaff/>),
- [2] Turing's 1950 paper <http://cogprints.ecs.soton.ac.uk/archive/00000499/00/turing.html>
- [3] For an online presentation (with Ron Chrisley) on ontological blindness in AI, robotics and psychology see <http://www.cs.bham.ac.uk/~axs/misc/talks/#talk16>
- [4] M.D. Hauser, 2001, *Wild Minds: What Animals Really Think*, Penguin Books.
- [5] The 'grand challenge' website:
http://umbriel.dcs.gla.ac.uk/NeSC/general/esi/events/Grand_Challenges/
- [6] More detailed child-robot project proposal: <http://www.cs.bham.ac.uk/research/cogaff/manip/>
- [7] Robocup Rescue: <http://www.r.cs.kobe-u.ac.jp/robocup-rescue/>
- [8] DARPA Cognitive Systems initiative <http://www.darpa.mil/ipto/> I shall attend their workshop 2-6 November.

Does this proposal meet the Grand Challenge criteria?

- It arises from scientific curiosity about the foundation, the nature and the limits of several scientific and philosophical disciplines.
- It gives scope for engineering ambition to build something that has never been seen before.
- It includes sub-challenges for which it will be obvious how far and when they have been met – others will take some research to specify.
- It could attract enthusiastic support from a wide multi-disciplinary research community, even those who do not participate and do not benefit from it.
- It has international scope: participation would increase the research profile of a nation (Cf. DARPA, Robocup rescue, the computer entertainments industry).
- It is generally comprehensible, and captures the imagination of the general public, as well as the esteem of scientists in other disciplines.
- It arises out of a challenge formulated long ago, which still stands.
- It promises to go beyond what is initially possible, and requires development of understanding, techniques and tools unknown at the start of the project, in particular new developments in computer science and software engineering, especially new theories about virtual machine architectures for intelligent systems.
- It calls for planned co-operation among identified research teams and communities.
- It encourages and benefits from competition among individuals and teams. though part of the task will be to develop criteria for evaluating progress (requirements).
- It decomposes into identified intermediate research goals, whose achievement brings scientific or economic benefit, even if the project as a whole fails e.g. many applications of AI, robotics, computer entertainments, and possibly computer-based education and therapy.
- It will lead to radical paradigm shift, breaking free from the dead hand of legacy, including forcing many narrowly focused researchers to start thinking about how their work can contribute to a much larger picture.
- It is not likely to be met simply from commercially motivated evolutionary advance.
- It requires cross-disciplinary collaboration including biology, neuroscience, social sciences, education, philosophy, linguistics, logic, mathematics, engineering, and all branches of AI.