

How To Think About Cognitive Systems: Requirements and Designs

Aaron Sloman¹

School of Computer Science, The University of Birmingham, UK
<http://www.cs.bham.ac.uk/~axs/>

Architectures have come to stay:

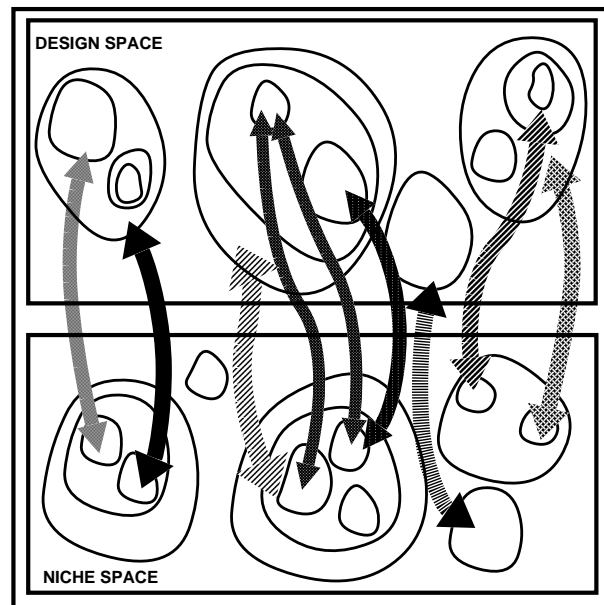
Much early thinking about AI was about forms of representation, the knowledge expressed, and the algorithms to operate on those representations. Later there was much in-fighting between factions promoting particular forms of representation and associated algorithms, e.g. neural computations, evolutionary algorithms, reactive behaviours, physics-inspired dynamical systems. More recently, attention has turned to ways of combining different mechanisms, formalisms and kinds of knowledge within a single multi-functional system, i.e. within one architecture. Minsky's *Society of Mind* was a major example.

So now people ask 'What architecture should we use?' My thesis is that this question is premature. We don't know enough to make grand choices, at least not as a community, though individuals can choose to investigate particular architectures. As a research community, we need more systematic investigation and comparison of a *variety* of architectures: including varied forms of representation, algorithms, mechanisms, communication links, and varied ways of combining them within a variety of architectures. We need to understand the advantages and disadvantages of different architectures in relation to particular kinds of tasks and task environments.

In other words, we need to study *design space* and *niche space* and the complex relationships between them.

The diagram on the right indicates that there are many interesting discontinuities in both spaces, and that instead of a single notion of "fitness" linking them there are complex (multi-dimensional) fitness relations determining trade-offs.

At present we understand little about the variety of possible designs for information-processing systems, especially designs for virtual machines, and perhaps even less about the variety of types of requirements: the tasks and constraints against which designs can be evaluated. In both cases we can learn from biological evolution.

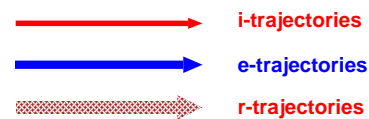
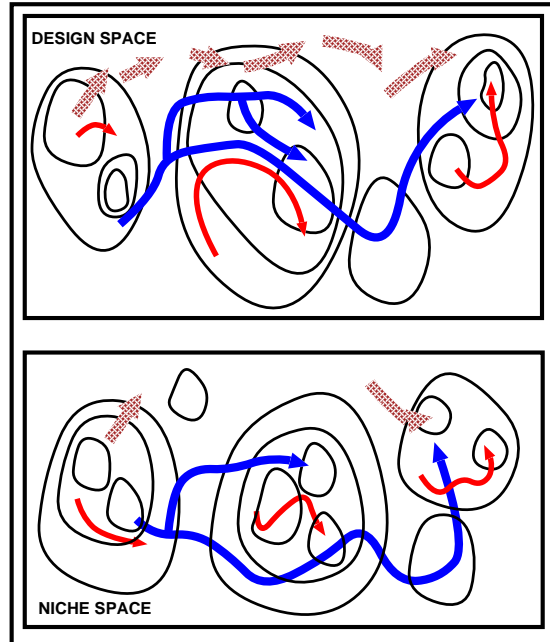


¹In collaboration with Ron Chrisley.

Trajectories in design space and niche space

Here are some types of trajectories in design space and niche space:

- I-TRAJECTORIES:
Individual learning and development
- E-TRAJECTORIES:
Evolutionary development, across generations, of a species or several species. These can be much longer with many branches.
- R-TRAJECTORIES:
Repair trajectories: An external agent intervenes. The process may temporarily disable the thing being repaired or modified.
- C-TRAJECTORIES:
E-trajectories where evolution is mediated by cognitive processes, e.g. mate selection.
- S-TRAJECTORIES:
Trajectories of social systems. Can be viewed as a subset of i-trajectories.



Members of *precocial* species have relatively short i-trajectories: they are born or hatched relatively well developed and competent e.g. chickens, horses. I.e. despite adaptation and learning most design information is genetically specified.

Members of *altricial* species are born or hatched relatively undeveloped and incompetent, e.g. eagles, lions, chimps, humans. They have long i-trajectories to allow for extensive bootstrapping (calibration, parameter setting, schema-instantiation, ontology construction?) during development, presumably because the information required by adults is too much or too variable to encode genetically.

During co-evolution of interacting species there are complex interacting feedback loops in both spaces. The same may be true for co-evolution of *components* (e.g. organs and competences) within a single species, e.g. co-evolution of perceptual capabilities, learning capabilities and deliberative capabilities. Understanding this may help us understand (and therefore help us replicate) complex products of such evolutionary processes, such as human minds.

We can learn from biological systems:

Myriad organisms process information of many kinds in many ways. These include:

- controlled production of individual organisms from eggs or seeds, and continuing growth and repair,
- evolutionary processes storing useful information for future organisms,
- complex emergent interactions among many relatively simple individual organisms such as termites building their cathedrals, or bees finding a good place for a new hive,
- most recently, relatively large, physically integrated, “expensive”, multi-functional minds and brains in individual organisms, e.g. in birds, cats, apes and humans,
- collections of individuals developing and transmitting a culture.

All of these processes are of great interest and are already being studied. However (d) is a particular challenge if we wish to understand and perhaps replicate aspects of human intelligence:

What are the requirements for building powerful, integrated, minds and what sorts of designs can meet those requirements? This has been a core goal of AI since its earliest days, but it is arguable that there has been very little progress relative to the overall goal, even though there has been much progress relative to what we knew fifty years ago.

Controllers and calculators, not Turing machines:

AI depends not on the idea of a Turing machine, as so many critics assume, but on merging two ancient strands of engineering:

- mechanical control mechanisms, e.g., speed governors, mechanical looms, music boxes, lathes, card sorters, etc.,
- machines operating on abstract entities, e.g., calculators of various kinds, machines operating on both numerical and non-numerical census information and eventually also machines that operate on their own programs [Ref1]. Physical machines operating on abstract entities were built before formal systems representing such entities and operations mathematically were available.

Electronic technology brought these two strands together in an ever-accelerating process of development of both 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. As always, science and engineering had to develop together. AI thus became the 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).²

The present context — almost total fragmentation:

Partly because AI has grown enormously, along with other relevant disciplines such as psychology and neuroscience, 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. Most (not all) vision researchers ignore work on natural language processing and *vice versa*. Robots are given various manipulative capabilities, but not the ability to understand or discuss those capabilities.[Ref3]

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. Perhaps it is now time for us to step back and ask:

- What types of machines are possible (including types of virtual machines)?
I.e. what is the structure of design space?
- What kinds of tasks can different types of machines do well or badly?
I.e. what is the structure of niche space, and how is it related to design space?
- What kinds of niches are suited to highly integrated multi-functional individuals, and what sorts of designs are appropriate?

²Some 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, unlike Turing, who modestly predicted in 1950[Ref2] “*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. Turing expected only a 30% success rate in a five minute test by 2000!

Addressing these questions requires us to expand our exploration of types of information-processing architectures and the variety of types of mental processes found in humans and other animals. In short, we need deeper, more systematic, exploration of the space of possible designs and the dual space of possible sets of requirements along with the variety of types of matches and trade-offs. (This can build on existing taxonomies of algorithms, taxonomies of neural nets, taxonomies of types of search, etc.)

We shall need extensions to the ontologies we now use for talking about designs and requirements, and new formalisms for expressing both. A great benefit of this work could be a framework for integrating much disparate research activity in AI, ethology[Ref4], neuroscience and psychology. But it will not be easy.

Difficulties – formulating requirements:

There is a subtle but huge obstacle that often goes unnoticed (a form of “ontological blindness”[Ref5]), namely the difficulty of understanding the *requirements* of the task. In particular we often underestimate the difficulty of discovering the capabilities of a child, a chimp, or a squirrel.

For example, many people think of perception as simply providing information about *physical* properties of the environment or about reliable correlations between *image patterns*. They therefore 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
- its capabilities
- the 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 its specific grasping capabilities – using teeth, fingers, or tail. Affordances will also depend on context (grasping while static or while moving, grasping with no obstacles or with intervening dangerous thorns, grasping from different directions, or while holding something else at the same time, etc.) A complex object such as a cup has (for adult humans) different affordances associated with its handle, its lip, the opening at the top, the sides, the base, etc. These affordances may change with context, for instance, which hand can more easily grasp the handle depends on the orientation of the cup and whether there are constricting objects nearby. There is no reason to believe a dog will see the same affordances, or a very young infant.

Since different parts of a complex object have different associated affordances, the affordances perceived in a complete object could be represented as *sets of condition-consequence relationships “attached” to structural representations of the various parts*. The conditions include possible actions of the perceiver and the consequences include possible outcomes that could be relevant to the perceiver. The sets of relationships are represented by being *attached* to appropriate components of structural representations of the object. This amounts to a richly structured collection of spatially-indexed counterfactual conditionals.³

This leaves open how the collections of conditionals are themselves represented. Two obvious options are sets of explicit rules with conditions checked against a symbolic database, and neural nets with associations implicit in connection strengths. Some people might consider that this should all be done using parametrised modal logics, though it somehow seems unlikely that animals naturally use modal logics, even if they are useful for some kinds of formal theorising about possibilities. What forms of representation are useful will in part depend on the ontology for states, processes, events and actions that the perceiver uses[Ref6] and its tasks.

³Exercise: consider what this implies regarding perception of a blank sheet of paper, for a child, a normal adult, or a Picasso.

Instead of just one solution, we may find different modes of representation of affordances useful for (a) the skilled and fluent performance of actions at high speed, and for (b) explicit reasoning about affordances, for instance in a deliberative mechanism. Yet another mode of representation may be required for (c) a reflective mechanism able to assess the implications of the information about affordances used by the deliberative system. Requirements may change in subtle ways if, besides being used privately by the organism, information about affordances is used publicly in giving explanations and advice to others, for instance an older child trying to get a younger child to understand why one way of stacking bricks makes them more stable than another way.

Conjecture: perceptual mechanisms⁴ in some animals have evolved so as to provide, in parallel, rapidly computed information about affordances at different levels of abstraction to different layers in a central sub-architecture. Contrast this with Marr's view of the functions of vision, namely to provide information about shape, motion, and surface properties such as colour [Ref7].

Conjecture: Since we appear to use spatial, and especially visual, competence in thinking about a wide range of non-spatial domains (e.g. search spaces), in many forms of problem-solving and communication, 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, until we understand human abilities to grasp and use visuo-spatial affordances[Ref9]. We have a long way to go.

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, learn to think about transfinite ordinals, discuss philosophical puzzles, and enjoy creating and experiencing poetry and string quartets.

One way to make progress is to take some of 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 preferences, motivation, emotions),
- (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 behaviours from design specifications. *This will probably require new formalisms for describing the complex architectures and the requirements that they have to meet.*

It would help if we could design a meta-theory for architectures: providing a way to systematically generate possible architectures covering a wide range of systems. (Compare languages and taxonomies for algorithms.) 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 suggest alternative architectures worth considering, possibly 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.

⁴Especially vision and hearing. This may be a precursor to fluent reading and fluent speech understanding.

An example:

The CogAff schema[Ref8, Ref10], described as an architecture in Rod Brooks' white paper on architectures for this workshop, is an incomplete first draft attempt to cover a wide variety of types of architectures, for humans, other animals, and possible future robots. It allows evolutionarily ancient reactive mechanisms to co-exist with and co-operate or compete with new mechanisms capable of doing different tasks, e.g., reasoning about what might happen, along with self-monitoring meta-management mechanisms of various kinds.

A special instance of CogAff, inspired by many facts about humans, is the H-CogAff architecture [Ref8] sketched on the right.

Goals can be generated in any part of the system (e.g. reactive hunger mechanisms and pain mechanisms, or ethical meta-management decisions). Some may directly (reactively) trigger internal or external actions, while others require deliberative resources in the upper layers. Variable, context sensitive, attention filters may reduce, but not eliminate, disruptive effects. The alarm mechanisms are part of the reactive system: i.e. fast and stupid, though trainable.

Using architecture-based concepts to define a mental ontology we can distinguish many kinds of affective states that can occur in H-Cogaff including several varieties of emotions.[Ref8] We can also account for many forms of arbitration and control, many forms of learning and development, etc.

This model appears to overlap substantially with the architecture independently proposed by a neuro-psychiatrist[Ref11]. There's also much overlap with Minsky's work[Ref12]. However, H-Cogaff still lacks many details, and further investigation may show a need to replace it.⁵

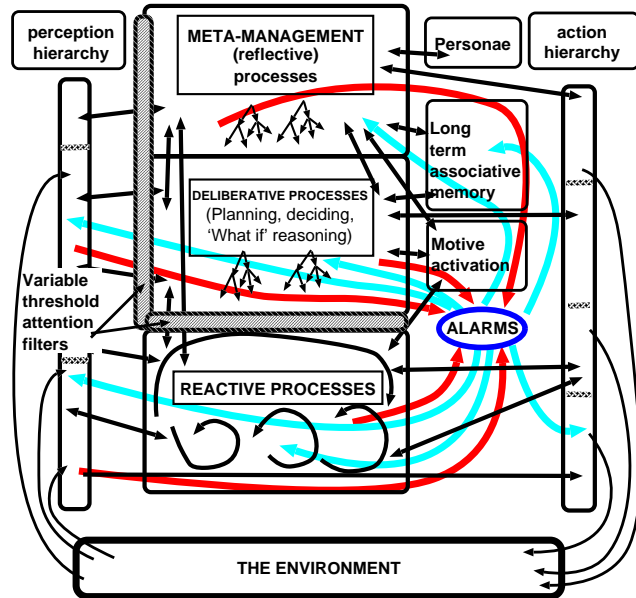
The CogAff framework allows us to consider possible i-trajectories involving H-Cogaff. E.g. it is unlikely that new-born infants have the full architecture. It is very likely that in humans and some other altricial species, architectural development occurs as part of the process of interacting with the environment. Similarly, this approach suggests that there are many more forms of learning than have so far been investigated, e.g. learning in different parts of the architecture.

Cogaff and H-Cogaff are merely examples. Better frameworks and architectures will require collaboration with neuroscientists and biologists, to help us discover information processing mechanisms not yet invented by engineers, and with psychologists and ethologists to find out in more detail what the requirements for complete systems are, in humans and other animals. Analysis of possible types of malfunction in proposed architectures could guide new empirical research on types of brain damage or disease and their consequences.

The proposal:

Given everything that has been learned 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 interacting robots, perhaps with the capabilities of somewhat simplified five year old children. There are (at least) three main tasks to be pursued in parallel:

⁵More information, including papers, slide presentations and software tools can be found at [Ref10].



- (a) Specifying *requirements* for the design to meet (which will include both (i) discovering what the requirements are and (ii) designing appropriate formalisms to express them.)
- (b) Producing a sequence of *designs* for robots to meet the requirements, including both designs for the physical body and designs for the virtual machine (mind). This may require extending current design formalisms and producing a richer generative architecture schema.
- (c) *Implementing* examples of such designs, some using physical simulations and some using advanced robot technology, and learning from limitations and failures.
- (d) *Abstracting*, if possible, a generic framework for adding “plug-ins” to allow a variety of humanoids to be produced.

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

Prospects:

Understanding how to build a child-like mind may provide a basis for many further developments. *Can we ever build a human child-like mind?* Not soon: we don't even understand the task.

We also do not yet understand the variety of mechanisms that can be built using all the computing power currently available: 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 producing a generative architecture-schema should help to accelerate progress in exploring specific architectures suited to different environments and tasks. This work has much potential for applications in robotics e.g. [Ref13], education, therapy, computer entertainments and many intelligent software-systems.

Some references:

- [Ref1] 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/>),
- [Ref2] Turing's 1950 paper <http://cogprints.ecs.soton.ac.uk/archive/00000499/00/turing.html>
- [Ref3] A draft proposal involving robots that can not only manipulate things but explain what they are doing, and help each other, is here: <http://www.cs.bham.ac.uk/research/cogaff/manip/>
- [Ref4] M.D. Hauser, 2001, *Wild Minds: What Animals Really Think*, Penguin Books.
- [Ref5] 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>
- [Ref6] A. Sloman (1996) Actual Possibilities, in *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*
- [Ref7] D. Marr, (1982) *Vision*
- [Ref8] A paper showing how at least three kinds of emotions arise in the H-Cogaff architecture is A. Sloman (2001), Beyond shallow models of emotion, in *Cognitive Processing: International Quarterly of Cognitive Science*, 2, 1, pp. 177-198,
- [Ref9] J. Glasgow et al. (ed) (1995) *Diagrammatic Reasoning: Computational and Cognitive Perspectives*, MIT Press
- [Ref10] Many more papers are available here <http://www.cs.bham.ac.uk/research/cogaff>
- [Ref11] R. A. Barkley (1997), *ADHD and the nature of self-control* The Guildford Press
- [Ref12] M.Minsky *The emotion machine* is on his website: <http://www.media.mit.edu/~minsky/>
- [Ref13] Robocup Rescue: <http://www.r.cs.kobe-u.ac.jp/robocup-rescue/>