Draft submission to Sunny Bains for AISBQ

How to build a human-like mind

Aaron Sloman
School of Computer Science, University of Birmingham
Birmingham, B15 2TT, UK
http://www.cs.bham.ac.uk/~axs

February 11, 2003

Many AI researchers nowadays focus their goals, their methods, and their (often unacknowledged) philosophical allegiances very narrowly. It's a way to make progress, get grants, get publications, and above all get demonstrable results.

For the next 300 years or so I intend to work on a different approach: trying to find out how to assemble a complete human-like mind.

The hardest part is coming up with a deep, broad and accurate account of what human minds are, or in other words, what sorts of things they can and cannot do. It is hard because most of what they do is invisible: external behaviour merely gives clues, and much of what goes on need not manifest itself externally. Neither can it be read off brain processes. Introspection helps but often leads to a mixture of excessive confidence and confusion.

We can make progress by collecting many different kinds of information from neuroscience, ethology, psychology, poetry, literature, philosophy, gossip, and even introspection and trying to weld it all into a coherent set of requirements for a *working* system.

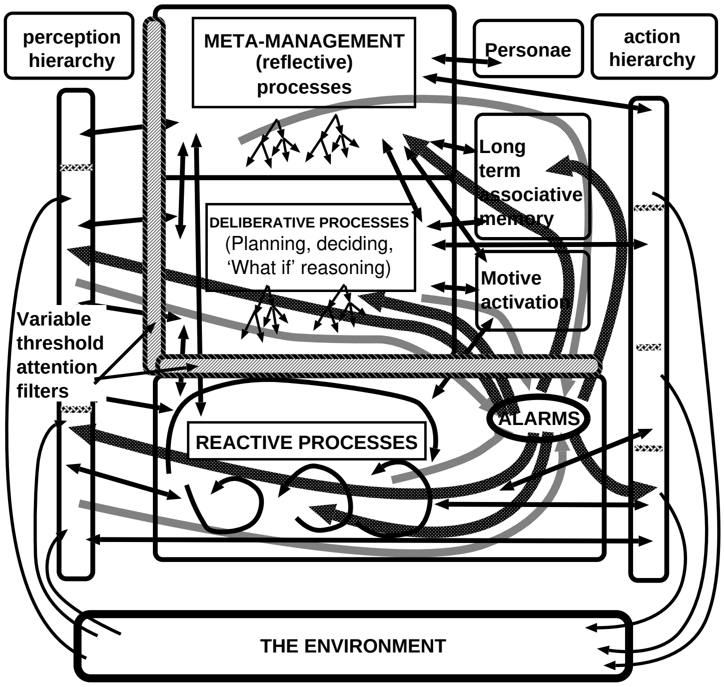
I have been doing that for about 30 years so far, and many of the results are available at the Birmingham Cogaff web site:
http://www.cs.bham.ac.uk/research/cogaff/
http://www.cs.bham.ac.uk/research/cogaff/misc/talks/

Alongside trying to come up with good characterisations of what we can and cannot do, I have been trying, with colleagues, to specify an information-processing architecture that can account for the phenomena. For that we need to have a good idea of what the space of possible architectures is, so that we can make an informed choice (instead of simply using the mechanisms favoured by our teachers, as often happens). Most of the interesting and powerful mechanisms are components of *virtual machine* architectures, and our understanding of the space of possible virtual machines and what they can and cannot do is in its infancy: it's such a huge space. So part of this project requires getting some sort of organised overview of that space.

We can do that partly by looking for different dimensions in which architectures, or parts of architectures can differ, taking account both of differences of function and also degree of sophistication, and likely evolutionary orders. Doing this led to a schema, the CogAff schema, for describing a class of architectures in terms of which sorts of components they contain and how those components are connected. This schema seems to be general enough to accommodate organisms as simple as insects (apparently using only reactive mechanisms that are mostly genetically programmed) many kinds of vertebrates, and also humans.

[Could include a diagram of the CogAff grid if there's space.]

A special case of the CogAff schema is the H-CogAff architecture, crudely depicted in the figure. It has many different components all operating in parallel with a wide variety of mutual influences, including hierarchical (layered) perceptual and motor subsystems and a variety of central mechanisms of varying evolutionary age, including reactive, deliberative and meta-management mechanisms, along with "alarm" mechanisms, various kinds of short term and long term memories, and a variety of types of forms of representations and ontologies deployed in the different layers.



Arrows represent information flow (including control signals). The alarm mechanisms receive information from many sources and using fast (but possibly stupid) pattern recognition mechanisms decide whether to send out high priority global signals to cause freezing, re-direction, or modulation of behaviours.

This architecture accounts for a wide variety of different sorts of perception, different sorts of action control, different sorts of learning and development, different sorts of emotions, different kinds of self-awareness, and different kinds of brain damage and their consequences.

In particular, if meta-management processes have access to intermediate perceptual databases, then this can support self-monitoring of sensory contents, leading robot philosophers with this sort of architecture to discover the problem(s) of "qualia". Some of them would become dualist philosophers.

It's a huge project: is it doable? I don't know. One way to make progress is to set up a long term target, and then identify a succession of increasingly difficult steps leading towards that target.

Such a target might be the design and implementation of a robot with a large subset of the abilities of a typical four or five year old child.

We can then attempt to achieve increasingly large subsets of those abilities. This contrasts with projects that aim to produce a neonate and let it learn: it is almost impossible to find out what is going on in new-born infants, and we know practically nothing about their mechanisms of learning and development.

Another option is to aim to produce something like a human adult. The problem with that is that adults contain huge amounts of culture-specific information and products of many years of individual learning, all of which can be very hard to identify.

By aiming for the capabilities of a young child who can talk, cooperate, argue, solve problems, explain things, etc. we may be able to identify a collection of capabilities that are generic to all cultures and provide the basis for a wide variety of different kinds of subsequent learning and development.

Such a project is outlined here:
  http://www.cs.bham.ac.uk/research/cogaff/gc

Collaborators welcome.