

**A grand challenge architecture project,
Re-combining various fragmented areas of AI, CS, and related disciplines**

**ARCHITECTURE OF BRAIN AND MIND:
HOW TO MAKE A HUMAN-LIKE ROBOT**

**Integrating High Level Cognitive
Processes with Brain Mechanisms and Functions**

**Notes, arising out of Panel D at the UKCRC meeting on grand challenges for
Computer Science, held in Edinburgh on 24-26th November 2002**

http://umbriel.dcs.gla.ac.uk/NeSC/general/esi/events/Grand_Challenges/

Last changed: December 19, 2002.

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

For some background ideas, see also

<http://www.cs.bham.ac.uk/research/cogaff/ibm02/>

<http://www.cs.bham.ac.uk/research/cogaff/manip/>

The project

**WE AIM TO BUILD A CHILD-LIKE ROBOT,
COMBINING MANY OF THE CAPABILITIES
OF A TYPICAL 4-5 YEAR OLD CHILD,
IN AN INTEGRATED FASHION, PROVIDING
A COHERENT PLATFORM FOR DEVELOPING A WIDE
RANGE OF HUMAN SKILLS AND KNOWLEDGE**

To be more precise: we'll design and build a succession of increasingly sophisticated such robots (physical and simulated), meeting carefully selected combinations of requirements to address deep scientific questions about the nature of brain and mind (and their relationship), while attempting to ensure that each step is both achievable and a major challenge.

Background

The idea of building something like a person has fascinated people for centuries:

E.g. the old “golem” idea, stories and films about man-made monsters (e.g. Frankenstein), likeable robots in science fiction (e.g. Star Wars, Forbidden Planet, AI), fearsome intelligent machines (e.g. COLOSSUS: The Forbin Project) and most recently AI-based synthetic agents in computer entertainments.

What most people find hard to appreciate, including many of the pioneering AI researchers, is how little we understand about what the requirements for a human-like system are.

Although we talk confidently about humans as seeing, thinking, learning, communicating, acting, being creative, having desires, intentions, feelings, emotions, and being conscious, we have no clear idea of what we mean, though we can point to many examples in everyday life.

A major feature of this project will be the combination of (a) analysis of **requirements** for satisfying these everyday descriptions of humans with (b) production and testing of **explanatory theories** and **implemented designs** for working systems.

The project will require a delicate balance between long term vision and practicality, if it is not to be as doomed as its precursors.

Architecture of Brain and Mind

- We aim to understand and model brain function at different levels of abstraction, including
 - Physiological properties of brain mechanisms, e.g. cortical microcircuits.
 - Neural information processing functions (probably requiring a new ontology of functions)
 - Higher level cognitive and affective functions of many sorts
 - Behaviours of complete agents (including social behaviours).
- This requires us to understand how the levels are linked to form an integrated functioning system (some levels implementing others),
- We aim to abstract **principles** of operation rather than mimicking biology in great detail.
- This requires us to
 - Identify important types of virtual machines found in natural intelligent systems
 - Identify important functional decompositions within virtual machines
 - Capture self-organisation and adaptability in addition to powerful specific capabilities and knowledge.
 - Attempt implementation-neutral specifications so that we can explore alternative implementations
 - Formalise ideas about requirements, architectures, representations, mechanisms, in mathematical theories (with help from theorists)
 - Develop working models, including both simulations and physical robots.
 - In particular, work towards building a working child-like robot

Also a grand challenge for computer science

This project will involve understanding and developing

- Coexisting virtual machines at different levels of abstraction, some implemented in others
- Functional decomposition into concurrently active subsystems at different levels
- Asynchronous interactions between sub-systems and between the whole system and its environment
- Simultaneous use of different kinds of formalisms with different sorts of semantics, E.g. in
 - low level vision
 - high level recognition processes
 - perception of affordances
 - planning and problem-solving
 - reactive mechanisms

Consequently this project generates a grand challenge to computer science to come up with formalisms, theories and tools for analysing, generating, and checking such systems.

This appears to be beyond the state of the art in computer science at present.

Compare the papers and slide presentations of Clark and Lam to the DARPA Cognitive Systems Workshop <http://www.dsic-web.net/meetings/oy8guwod/presentations.html>

Progressive-deepening strategy - (1)

Because the task is so horrendously complex, success will depend on very careful selection of intermediate goals, that are (a) achievable and (b) launching pads for further work.

- Trying to simulate a **complete adult human**, even at a high level of abstraction, would be impossibly difficult because of the huge amount and great diversity of knowledge required, and could be ill-advised because there are so many differences between adult humans that any such simulation would largely reflect ad-hoc individual history and cultural influence, probably obscuring deep general principles. (Likewise “intelligent assistant” projects.)
- At the other extreme, simulating a **new-born infant** would raise many methodological problems, including both identifying the internal processes of a new-born child and also producing behaviour of a type that could clearly demonstrate that we are going beyond superficial behavioural manifestations (gurgling, crying, waving arms and legs, sucking, eye-movements, etc.)
- We therefore propose aiming to simulate a child, aged between 3 and 5, able to perform many tasks that reveal different competences, including the task of communicating with others, asking for and giving advice and help, and purposeful manipulation of physical objects. We shall identify a level of competence that could be a basis for a wide variety of types of human learning.
- Neural theories will mainly provide general constraints rather than complete implementations of the neural infrastructure.

Progressive-deepening strategy - (2)

An early part of the project will involve

- Analysing requirements for modelling a complete human-like child of the specified age.
 - We already have very broad but shallow knowledge of the requirements, e.g. much common knowledge about human capabilities, enhanced with decades of research in psychology (e.g. observations of Piaget, Vygotsky and many other developmental psychologists)
 - Many details are far more obscure than most people realise, e.g. *what do we see when looking at a piece of machinery or at a block structure to be copied, or drawn or described?* Something like the notion of **affordance**, seems crucial, not just objects and relationships.
- Identifying important core subsets of the requirements which could be put together in complete working models able to demonstrate their competence.
- Finding an initial *sequence* of such models to aim at that could be
 - Achievable (albeit with great difficulty)
 - Demonstrable in a succession of working (physical or simulated) robots
 - Each providing a launch pad for achieving the next model in the sequence
 - Leading to a robot demonstration in 10 or 15 years that is far beyond the state of the art at present and clearly on the way to a human-like child that is capable of developing further and learning the sorts of things that produce adult humans in different cultures.
- Identifying major gaps in our knowledge, understanding, techniques and tools; and initiating projects to fill those gaps.
(Some results may not be deployable until later models are constructed.)

Progressive-deepening strategy - (3)

Progressive deepening can have two concurrent theoretical strands:

- **The project includes**
 - (a) **people who have already spent many years analysing “top-down” and “middle-out” the requirements for human-like systems and**
 - (b) **people who have spent many years investigating natural and artificial neural mechanisms and other brain mechanisms.**
 - **Together they can provide an initial architectural framework for systems that combine reactive, deliberative and reflective mechanisms that might be implemented in a variety of types of neural mechanisms, along with hierarchically organised concurrent perception and action mechanisms (including language understanding and production).**
- **Others have focused on specific human (and animal) capabilities.**
 - **This includes perception, learning (including ontology formation), memory, reasoning, understanding spatial structure and motion, understanding causation, thinking about and interacting with other intelligent agents, problem solving, planning, plan execution, motor control, dealing with the unexpected, affective states and processes (e.g. desires, emotions, preferences) and use of natural language and other forms of external representation.**

The specific results of the second sort of research may have to be modified to slot into the general architectural framework produced by the first sort. And the architecture may need modification to accommodate some of them.

At first, simplified forms will be combined, then gradually richer and more complex versions, and more of them.

Progressive-deepening strategy - (4)

Besides the progressive deepening of **theory** we should aim for a succession of deeper and broader Robot-Child **behavioural specifications**.

- Some small steps towards a target system about 3 and 5 years hence can be found in this incomplete draft project specification (in postscript and PDF):

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

(Based partly on discussions with Marvin Minsky and Push Singh, MIT.)

- A subset of that specification could be selected for the first two year target, involving very great simplification of the physical design of the robot so as to retain many of the deep problems while preventing complete intractability.

The early robots might not have a very engaging (humanoid) appearance, provided that their sensory and motor capabilities pose a deep enough range of problems, e.g. perception of affordances related to the robot's manipulative capabilities and goals.

- As the work progresses it will be tested on a succession of increasingly sophisticated robots, both physical and simulated, as in the international RoboCup and RoboCupRescue projects.

<http://www.r.cs.kobe-u.ac.jp/robocup-rescue/>

- We could try to set up an international RoboChild project in parallel with those.
- There may be a related project funded by the DARPA Cognitive Systems programme, on which I have made some comments here:

<http://www.cs.bham.ac.uk/research/cogaff/darpa02/>

Research to build on

Existing work this project can build on includes:

- John McCarthy's work on the well-designed child (summarised below):
- Marvin Minsky's draft chapters for *The Emotion Machine*
<http://www.media.mit.edu/~minsky/>
- Work in the Birmingham Cognition and Affect project
<http://www.cs.bham.ac.uk/research/cogaff/>
- Stan Franklin's models based on B.Baars' theory of consciousness.
- Michael Arbib's work on 'Conceptual neural evolution' presented at the WGW02 conference August 2002.
- A vast amount of more specialised research in
 - AI (including robotics, NLP, learning, vision, connectionism, etc.),
 - neuroscience,
 - psychology ((e.g. recent work on “executive functions” by Barkley and many others.)
 - biology (including ethology, e.g. M. Hauser's *Wild Minds*)
 - linguistics,
 - language and toolkit design
 - etc.,

NB: there have been huge advances in robot capabilities in recent years, but not in giving the robots the ability to understand what they are doing, and to explain it to others and discuss rival options. Vision remains mostly low level.

McCarthy's Well Designed Child

John McCarthy's paper "The Well Designed Child" lists important features of our environment, and some capabilities required to cope with them. See <http://www-formal.stanford.edu/jmc/child1.html>

He discusses kinds of innate knowledge and abilities, concerning:

- the existence of persistent objects, forming natural kinds, with 3-D structure, locations, colours, relations, continuous motion...,
- kinds of situations, some of which recur,
- curiosity focused on information likely to be useful, treating other things as "probably noise",
- goal/sub-goal hierarchies and "the grammar of goal regression",
- the need for introspection, and the developing ability to do it,
- the "principle of mediocrity": I can learn about myself from observing others and vice versa,
- various kinds of meaning (to which grammar is secondary),
- the ability to process some kinds of information in parallel,
- the roles of logical and non-logical forms of representation (including use of chemical states to represent biological needs) – some of these express "virtual sentences",
- how to compose whole thoughts from components,
- differences in linguistic requirements for perceiving, thinking (including use of pointers) and communicating, (Includes thinking about future possibilities.)
- varieties of reasoning using different forms of representations, including parallel inferences.

We can test McCarthy's proposals by finding out whether they are implementable and how well they work.

Beware of the “emotions” fad

Many authors assume that emotions are needed for intelligence (following work of Damasio and others). However, the study of emotions is full of confusions:

- Most emotion theorists don't know how to explain states in terms of what's going on in an information-processing architecture.
- So they try to characterise emotions in terms of observable/measurable phenomena (blood pressure, galvanic skin response, weeping, smiling, posture, tone of voice, etc.)
- Instead we should develop an **affective ontology** based on varieties of control processes within an architecture, e.g. generation of motives, detection and resolution of conflicts of motives, priority changes, alarm mechanisms and interrupts of various sorts, switches of attention and modes of processing, etc.
- CogAff papers show how to a first approximation, at least three major categories of emotions (primary, secondary, tertiary emotions), are associated with the three main types of architectural layers (reactive, deliberative, reflective).
- **A full theory would start from our common-sense taxonomy of desires, preferences, pleasures, pains, values, ideals, attitudes, concerns, interests, moods, emotions, intentions, etc. and then produce a richer, more precise, architecture-based taxonomy of affective control states.**

Compare: H.A. Simon 'Motivational and emotional controls of cognition', 1967.

Some examples of work to be included

- The position paper by Mark Steedman pointed out that systems uniting high level structural modes of analysis and low level statistical modes of analysis can already outperform serial systems of either sort. This is a natural consequence of the multi-layer architecture. Examples include:
 - Major improvements in wide-coverage parsing and speech recognition
 - The return of high level vision in combination with lower level approaches
 - Hybrid problem solvers, planners, theorem provers, combining learnt patterns with general reasoning powers, etc.
- Work on bootstrapping structural representations from neural representations
- Attempts to provide a better understanding of low level brain processes in the light of their support for higher level virtual machines. This could lead to radically new computing paradigms and new questions for neuroscience to answer.
- Attempts to design new brain-inspired hardware (as in Steve Furber's proposal),
- Novel approaches to self-organisation and adaptation.
- Perceiving affordances and reasoning about them affordances,
- Self monitoring, self evaluation, self control.
- Affective processes (mentioned previously),
- Understanding others: empathetic and simulative reasoning; dialogues.
- Applying models of creativity to creative processes in a child such as modifying affordances (e.g. finding a new way to prize a lid off a tin of blocks).

Potential Applications

The project has a very wide range of potential applications, some immediate and some requiring further developments based on the results.

- **Many robotics applications, including**
 - Aids for aged and infirm (allowing elderly people to live at home longer, or to be less dependent on human help).
 - Intelligent guides for the blind (more flexible than dogs?)
 - More intelligent robots in factories - able to learn and cope with surprises,
 - Robots in hostile environments acting autonomously, and communicating verbally instead of having to transmit images for remote human controllers.
 - New kinds of “teaching robots”
 - and many more ...
- **Education based on improved understanding of how learners work**
- **Intelligent assistants and embedded intelligence of many kinds**
- **New ideas to help advance brain science and psychology (e.g. providing better explanations for multiple visual pathways, or for attentional disorders),**
- **Many applications in entertainments industries.**
(Potential multi-billion pound payoff.)

Criteria for success

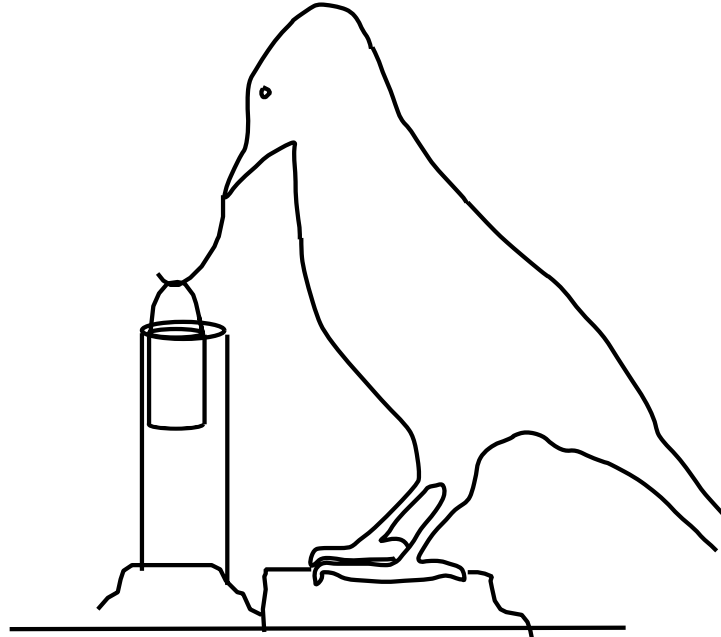
There is no achievable “final” stage attainable in the foreseeable future. But there are many worthwhile intermediate stages that provide grand challenges.

Tests for success could take several different forms:

- Demonstrating increasingly difficult robot capabilities: A convincing child-like robot (<http://www.cs.bham.ac.uk/research/cogaff/manip/>) that can
 - act in an everyday environment
 - explain its actions
 - help others improve theirs
 - learn new skills and concepts etc., including expanding linguistic competence,
 - talk about and reason about the mental states of others
 - develop its understanding of numbers over time
 - learn to prove, or even discover, simple geometrical theorems
- Demonstrably improve theories in psychology, linguistics and education
- Lead to important new therapies and counselling techniques for some mental disorders (Cf. R.H.Barkley on ADHD)
- Inspire significant new research in neuroscience, e.g. posing new questions for brain imagers to investigate, e.g. about nature and causes of autism.
- Demonstrably improve attempts to understand how other animals perceive, learn, solve problems, etc
- Development of many new kinds of embedded intelligent systems and intelligent software aids and companions.

Betty Crow: Cognitive Agent and Hook-maker

Two crows, Betty and Abel, learnt to use bent wire to fish a bucket of food out of the vertical tube (as in the picture). Then Abel flew off with the hook.



See the video here: <http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm>

To find more, give google: **betty crow hook**

- Betty tried using a straight piece of wire for a while, and failed.
- She then pushed one end of the wire into the tape holding the tube and moved the other round with her beak, making a hook, which she used to lift the bucket.
- She did this 9 times out of 10. **Reported in Nature and shown on BBC TV (August 2002).**

COULD OUR ROBOT REPLICATE BETTY'S MENTAL PROCESSES?

What sort of architecture could do what Betty did?

Many explanations are compatible with **any** observed performance, e.g.:

- **Pure chance?**
- **An innate behaviour** triggered by some mixture of internal and external state?
 - What mixture?
 - How did the genes get the information? Why was it selected?
- **A learnt adaptation** in a trainable (altricial) reactive system?
 - What sort of boot-strapping could achieve this?
 - How is the learnt information acquired, represented, stored, activated, used?
- **Was it a deliberative** (e.g. problem-solving) process?
 - Using what sort of ontology for possible goals, states, actions?
 - Using what general knowledge?
 - Invoked how?
 - Acquired how? (Using an architecture built in infancy?)
 - Using what planning mechanisms? (Using what representations, what search mechanisms?)
- **Did it involve self-knowledge?** (Reflection/meta-management)
 - Did Betty understand what she was doing, or did she, like many AI deliberative systems, lack reflection/meta-management? (Can a crow teach another crow to do this?)

The questions are deep and important because understanding of spatio-temporal processes can be re-used in many contexts.

E.g. doing mathematics, designing architectures, thinking about anything complex.

Vision and affordances

Vision is not just about

- Object recognition
- Perception of geometrical and physical structure and motion
- Building cognitive maps for route-planning

There's something deeper, not yet properly characterised, which can be called **perception of affordances**.

- Affordances are not “objective” properties intrinsic to physical configurations.
- They are **relational** features dependent on the perceiver's
 - Common or likely goals and needs
 - Capabilities for action (physical design + software)
 - Constraints and preferences (avoid stress, injury)

What affordances did Betty need to see?

What sort of robot child could see them, and use them to solve similar problems?

Towards a theory of affordances

Affordances in a complex scene can be construed as

- **sets of sets** of counterfactual conditionals,
- **spatially indexed**: different sets attached to different parts of objects.

How should affordances be perceived, represented, used, explained to others?

Different representations and mechanisms handle affordances in different architectural layers -

e.g. skilled behaviour is mostly reactively controlled.

Probably not using modal logics???

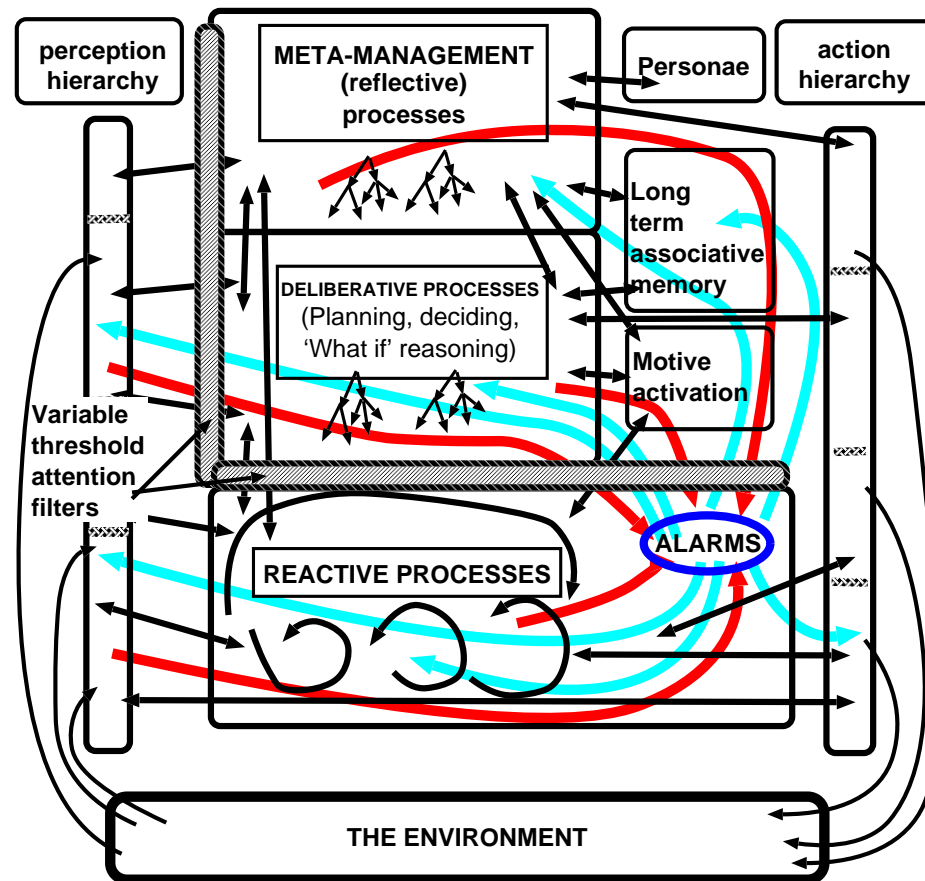
Can we get some hints by exploring possible neural under-pinnings for sets of conditionals stored as context-sensitive associations?

This still leaves many hard questions about short term memory.

A hypothetical Human-like architecture:

H-CogAff (See <http://www.cs.bham.ac.uk/research/cogaff/>)

This partly overlaps with Minsky's *Emotion machine* architecture.



Where could it come from?

Various trajectories: evolutionary, developmental, adaptive, learning...

TOOLS - A PUSH-PULL RELATIONSHIP

A project like this will require powerful tools of many kinds. As tools are developed they will help with progress in the project. As new problems are discovered this will identify requirements for new and improved tools. The project may help to drive integration of the following:

- Many agent toolkits. Some support **specific** agent architectures (e.g. SOAR), while others are **architecture-neutral** (e.g. the Birmingham SimAgent toolkit). Some are primarily aimed at building complex **individual** agents while some are more concerned with interacting **collections** of relatively simple agents. Some try to do both by treating a complex agent as a collection of simple agents. (But will a *homogeneous* collection suffice?)
- Games toolkits. Often these provide very fancy high speed graphics but practically no support for putting “intelligence” into the agents.
- Physical simulation engines. These are being used increasingly in games and movies that depend on very realistic displays of complex physical simulations (e.g. the film Titanic.) They do not aim to support development of mental processes.
- Tools for building “humanoid avatars” for synthetic worlds. These provide mechanisms for biologically plausible shapes and forms of motion, including, in some cases facial expressions, gestures, etc. However typically they ignore the problem of modelling cognitive and affective processes.

There's a lot more to be said

ENOUGH FOR NOW

This is an invitation to researchers in CS, AI, and many relevant disciplines to join this project.

Acknowledgements

My thanks to members of Panel D at the Grand Challenges workshop,
Johanna Moore (panel chair)

Mike Denham (Now the project “champion”)

Mark Steedman

Simon Colton

John Sutherland

Leslie Smith

And others who contributed papers but could not attend.