DRAFT: likely to be updated

Machines in the ghost

Aaron Sloman

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

Based on work done with Jackie Chappell,
Dean Petters, and the Birmingham CoSy team
on requirements for human-like robots.

NOTE:
I try to make my slides readable by anyone interested – without having to hear me present them.
This means

(a) that they contain too much clutter for presentations
(b) that there are usually far more slides than can be included in a single presentation.

NOTE: these slides are produced using LaTex and developed and presented on linux.

Abstract (revised slightly after conference)

- This paper summarises a subset of the ideas I have been working on over the last 35 years or so, about relations between the study of natural minds and the design of artificial minds, and the requirements for both sorts of minds.
- The key idea is that natural minds are information-processing machines produced by evolution.
- What sort of information-processing machine a human mind is requires much detailed investigation of the many kinds of things minds can do.
- In particular, it is not clear whether producing artificial minds with similar powers will require new kinds of computing machinery or merely much faster and bigger computers than we have now.
- Insofar as some sorts of psychotherapy (including psychoanalysis) are analogous to run-time debugging of a virtual machine, in order to do them well we need to understand the architecture of the machine well enough to know what sorts of bugs can develop and which ones can be removed, or have their impact reduced, and how.
- Otherwise treatment will be a hit-and-miss affair.
- This requires understanding how minds work when they don't need therapy – a distant goal.

The online conference programme is at `http://www.indin2007.org/enf/program_overview.php`
The full paper (Machines in the ghost) is available here

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0702`

There are three (overlapping) kinds of machine

1. Matter manipulating machines
2. Energy manipulating machines

   Many kinds, of both sorts, have been investigated, designed, and built, by humans for thousands of years.

   Before that, such machines were produced by evolution: e.g. body-parts of animals, many still unequalled by any man-made machines.

3. Information-manipulating machines

   With many sub-divisions, e.g.

   - Passive: requiring energy to be constantly supplied from outside, like an abacus or mechanical loom
   - Active: with a store of deployable energy, and the ability to renew it when necessary (by using some of it to obtain more 'fuel' from the environment).

   WITH VARYING DEGREES AND KINDS OF MODIFIABILITY BETWEEN THESE TWO EXTREMES:

   - Fixed, with an unchanging repertoire of capabilities
   - Self-assembling, self-extending, self-repairing

   Parameter adjustment within a fixed structure is a very simple kind of self-modifiability: contrast structure-modification, e.g.

   - acquiring the ability to understand Urdu
   - acquiring the ability to sight-read piano music
   - acquiring the ability to think about transfinite ordinals and the ability to enjoy it
   - acquiring the ability to design and debug computer programs.

Varieties of information-processing

Information-processing is not just a collection of syntactic operations, e.g.:

- bit-manipulation,
- storing
- retrieving
- pattern matching,
- transfer of data from one place to another.

A crucial feature of an information-processor is CONTROL

If a machine cannot do anything then there is no sense in which it understands any of the information it contains, however useful the information is for something else or someone else, e.g. information we store in a filing cabinet means nothing to the cabinet.

Moreover, in an organism or machine with mere cognitive competence (e.g. sensing mechanisms, mechanisms for deriving consequences from information, planning mechanisms, hypothesis-forming mechanisms) nothing would happen:

without desires, preferences, values,... there would merely be unrealised potential for many happenings: nothing would be initiated.

DAVID HUME: "Reason is and ought to be the slave of the passions and can never pretend to any other office than to serve and obey them."

(A misleading over-statement: it's a mutual relation.)

Some varieties of control

Control can be

- continuous or discrete,
- sequential or concurrent,
- either occurrent (sub-system actually controls processing)
  or dispositional (sub-system attempts to control processing but is over-ridden or suppressed by others, until circumstances change.).

and effects of control mechanisms can include

- initiating processes
- modulating, suspending, aborting processes
- considering and evaluating options
- inventing new alternatives
- evaluating new internal or external sensor data
  evaluations can be triggers for new control processes
- detecting conflicts
- resolving conflicts
- competing sub-processes
- collaborating sub-processes

Show two demos – simulated “emotional” agents and reactive and deliberative sheepdog herding sheep.

More varieties of control

In a complex multi-function control system there may be many different kinds and levels of sub-systems, which can vary in all the ways previously listed, which may be:

- concurrently active,
- self-monitoring,
- self-modulating,
- self-controlling

I have argued (starting in The Computer Revolution in Philosophy (1978)) that many of our ideas of consciousness are really ideas of self-consciousness

especially the kind of self-consciousness that would arise in a very high level globally monitoring and controlling sub-system – required because other mechanisms for resolving conflicts, like voting are inadequate (e.g. because they can be too stupid!)

Contrast Baars' Global Workspace theory

Many of the things Panksepp, Solms and others say about consciousness and feelings are related to varieties of self monitoring and self control, which involve an architectural layer that is not absolutely required for the functioning of the rest of the system,

which is why some organisms (e.g. all insects?) function without them.

but which, if present, bring certain benefits (and costs).

Some control engineers have begun to appreciate the need for this (e.g. Ricardo Sanz, Dietmar Dietrich).
Also John McCarthy "Making Robots Conscious of their Mental States" (1999 version)

Minds are ....

1. Active,
2. self-assembling, self-extending, self-repairing
3. structure-modifying
4. self-monitoring, self-modulating, self-controlling
5. information-processing
6. virtual machines.

The need for a great deal of variability in a mental machine implemented on a physical machine with a fixed structure, requires the mental machine to be a virtual machine.

It also requires the physical machine to have sufficient ability to switch between different states to meet the needs of the virtual machine: the PRINCIPLE OF SUFFICIENT VARIABILITY.

In brains and computers a fixed global structure is combined with vast amounts of local switching.

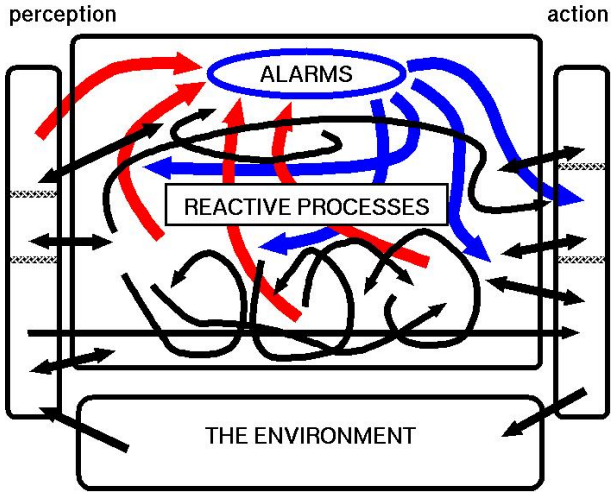
The requirement for concurrency in mental virtual machines adds additional requirements for the physical implementation machines.

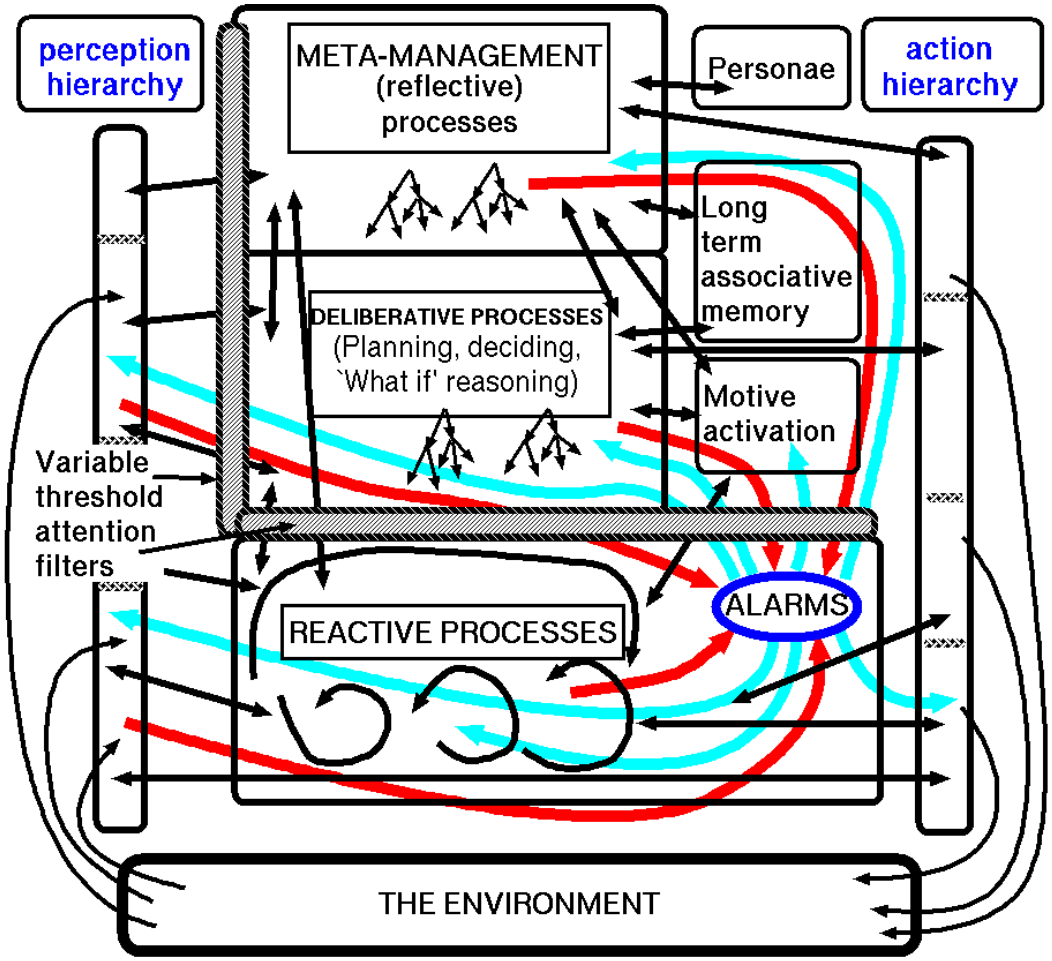
Concurrency can be implemented directly or indirectly

e.g. indirectly through time-shared processing streams,

with different consequences that require more analysis than I have time for.

Biological virtual machines vary enormously





Reactive mechanisms (sketched above) include reflexes (direct links from sensors to effectors) and many other things.

They exclude representing hypothesised past, remote or future situations, and cannot build multi-step plans.

But their perceptual and motor subsystems may use different ontological layers.

Between purely reactive microbe-like or insect-like machines (architecture on left) and vast ecosystems, there are multi-level self-extending human-like virtual machines (crudely depicted on right)

(See `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`)

How to think about minds: be designers

- Minds DO things:

  they learn, perceive, decide, remember, imagine, plan, forget, grieve, enjoy, experience itches, want to scratch, prove theorems, seek admiration, feel guilty, etc.
- Any kind of doing requires some kind of machine.
- Doing what minds do requires at least an information-processing machine.
- We understand very little about information processing machines: we have only recently begun to study them by designing, building, debugging, comparing, various kinds including self-modifying machines of various sorts.

But all current artificial machines lack the vast majority of features of even a young child.

A child that has not yet learnt to talk has a deeper understanding of possible things to talk about than any AI system built so far. That's what drives the language learning process.

But that's another long story, see

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0702`
Sloman & Chappell: Computational Cognitive Epigenetics (BBS 2007)

We need to test our theories by designing, implementing, testing and debugging working systems: armchair designs usually cannot work.

See `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/design-based-approach.html`

How a machine experiences things

There is much talk in philosophical and psychoanalytic (and other) circles about having a point of view, as something that resists explanation in information-processing terms. We can analyse this as having two components:

- what a machine senses has a (changeable) point of view which may be unique to that machine
- what's going on in a machine can be sensed and recorded and used by the machine if its architecture has a meta-management component.

If a surveillance camera is mounted on apparatus that moves it round and makes other changes, what the camera records changes according to its exact position and orientation, and perhaps other things such as current depth of focus, whether specific optical filters are used, etc.

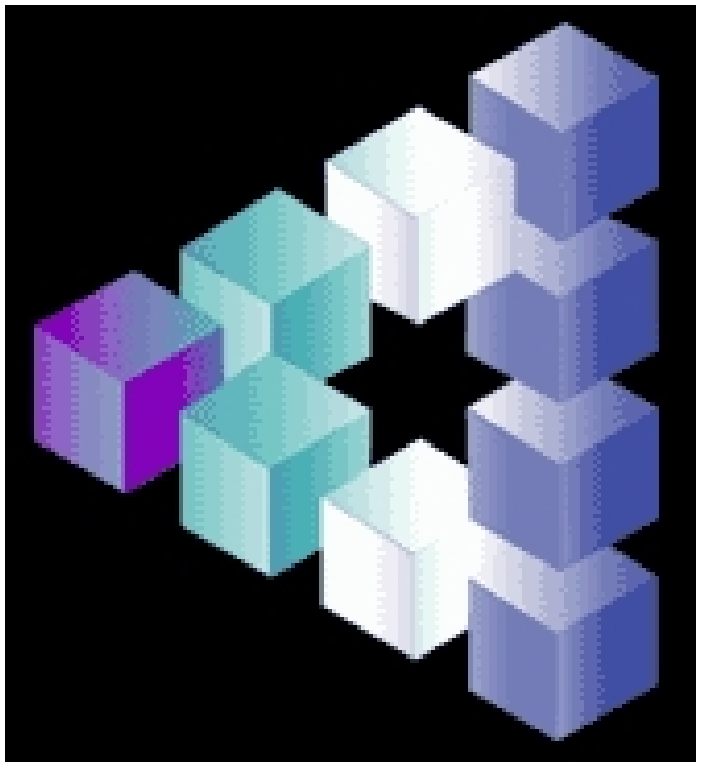
For instance, which objects and parts of objects are seen, where things are in the camera's 'visual field', what the aspect ratios are, etc., can change.

- Every animal and every sensing device has a 'personal' view of the universe. I.e. what it is like to be that machine or animal is different from what it is like to be any other.
- Some organisms, in addition, have a reflective, or meta-management component, which allows an animal (or machine) to detect, record, and use information about what its point of view is at any time.

What evolution can teach us

- Evolution discovered and built many more kinds of information processing machines, on many more scales of complexity than we have studied.
- In doing so evolution solved many more kinds of problem than we have so far understood:
- so one of our main tasks is to find out what the problems were
- The problems evolution solved are not obvious – e.g. we are not yet able to characterise the requirements for human vision properly!

Example: What set of requirements led to a virtual machine that can be driven by sensory data to construct a percept of an impossible object?



High level percepts can be inconsistent
(Picture by Reutersvard – before Penrose)

This tells us important things about the visual system – and some of the contents of visual consciousness.

What you see is not only what exists, but also multiple affordances.

Think of all the things you can do with or between the little cubes.
Collections of affordances can be inconsistent: but not models of a scene.
If the picture were huge, you might never discover the impossibility (like a 2 year old).

Compare Escher's pictures, e.g. the Waterfall.

And how can a virtual machine that that doesn't detect the impossibility (e.g. a two year old human?) change itself into one that does (e.g. a teenager, or mathematics professor)?

See also `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#compmod07`

Multiple perceptual routes

H-CogAff specifies multi-window perception and multi-window action,

The visual and action sub-systems have architectural layers (evolved or developed) that

- concurrently handle ontologies at different levels of abstraction (including in some cases mental states of oneself and others),
- and have multiple connections to different sorts of central sub-systems, as well as to other sensory and motor subsystems.

In contrast many architectures assume peephole perception and action.

I.e. they assume that perceptual sub-systems process only low-level image data.

That would mean that musical sight-reading or a tennis player's visual expertise is not part of the perceptual (visual) system.

So, instead of one or two routes from vision, we have multiple routes,

e.g. to blinking reflexes, saccade generators, posture control subsystems, visual servoing mechanisms, motive generators, question answering mechanisms, planning mechanisms, prediction mechanisms, explanation constructors, plan execution mechanisms, learning mechanisms (in several different architectural layers), alarm subsystems, communication mechanisms, social mechanisms.

SIMILAR COMMENTS APPLY TO CONNECTIONS WITH ACTION SUB-SYSTEMS.

Common-sense concepts can obstruct research

When scientists and engineers discuss what needs to be explained, or modelled, and when they report experimental observations, or propose explanatory theories, they often use concepts that evolved for use in informal discourse among people engaged in every day social interaction, like this:

- What does the infant/child/adult/chimp/crow (etc) perceive/understand/learn/intend (etc)?
- What is he/she/it conscious of?
- What does he/she/it experience/enjoy/desire?
- What is he/she/it attending to?
- What sort of emotion is he/she/it having now?

In ordinary discourse, plays, novels and even in law courts, these words work fine.

And they can be pointers to empirical phenomena that need research in order to find evolutionary origins, neural mechanisms, individual differences.

But if we treat them as precisely defined theoretical concepts we can end up asking non-questions and exploring theories that are too unclear to have explanatory value.

What alternative is there?

Yet we cannot avoid using them initially

As Jaak Panksepp says in his commentary on my paper: we have no choice but to start from ordinary concepts, in a long term scientific bootstrapping process

as also happened in the physical sciences: physics, chemistry, geology, astronomy ...

But we can also be mistaken about how our ordinary concepts actually work:

See The computer revolution in in philosophy (1978) Chapter 4 on conceptual analysis.

In particular, it is common (but unfortunate) that the word 'emotion' is used to refer to things that are highly disparate, for which ordinary language has many subtle and distinct labels

including: desire, dislike, preference, enjoyment, goal, intention, idle wish, inclination, irritation, anger, indignation, rage, having values (including moral and aesthetic values), attitude, ideal, obsession, jealousy, embarrassment, awe, shyness, happiness, sadness, mood, disposition, schadenfreude, thirst, lust, sexual enjoyment, feeling things (ordinarily used in many different ways), finding interesting, remorse, regret, guilt, pity, self-pity, and many more

Novelists, playwrights, poets and garden-gate gossips have much implicit knowledge about these matters, which they use successfully in many activities including both private and social activities.

But for the activities of scientific explanation, and the associated forms of engineering, including medicine, psychotherapy, etc., we need much deeper, theory-based concepts: where a theory includes a design for a working system (roughly).

The design-based approach

We are talking about complex systems with many concurrently active parts, doing many different things, that work together more or less harmoniously most of the time but can sometimes come into conflict.

These parts are organised in an information-processing architecture that maps onto brain mechanisms in complex, indirect ways that are not well understood.

So we should ask questions like this if we wish to do deep science:

- Which parts of the architecture are involved?
- What are their functions?
- What kinds of information do they acquire and use?
- How do they do this?
- What is the total architecture in which they function?
- How is the information represented?
  (It could be represented differently in different sub-systems).
- What kinds of manipulations and uses of the information occur?
- What mechanisms make those processes possible?
- How are the internal and external behaviours selected/controlled/modulated/coordinated?
- How many different virtual machine levels are involved and how are they related (e.g. physical, chemical, neural, sub-symbolic, symbolic, cognitive,...)?

`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/design-based-approach.html`

What is information-processing?

- Researchers in many different disciplines talk about information and information processing, but very few understand what they are talking about, because they have never built, debugged, extended, working information processing systems.
- Software engineers have a deep practical understanding of the issues, but don't know how to articulate it: they are not trained in philosophical analysis.
- Most scientists (e.g. neuroscientists) who talk about information processing don't realise that the most important information processing machines are virtual machines whose structure and function cannot be read of physical machines.
- Philosophers unwittingly make hugely oversimplifying assumptions about information processing, e.g. assuming that if you know about Turing Machines you know what information processing is.

  They often assume that virtual machine events cannot be causes: they must be epiphenomenal – software engineers know that is false, as many bugs are effects of virtual machine events.

  There is also often confusion between virtual machines and programs:
  A virtual machine does not exist just because a program exits: the program must be running.
- 'Information' cannot be defined explicitly without circularity, any more than 'matter' and 'energy' can: but all three are defined implicitly by the theories in which the concepts are used. (A technical notion that needs more time than I have.)
- There are many more types of biological virtual machine than anyone knows about – including virtual machines implemented in chemical (molecular) processes.

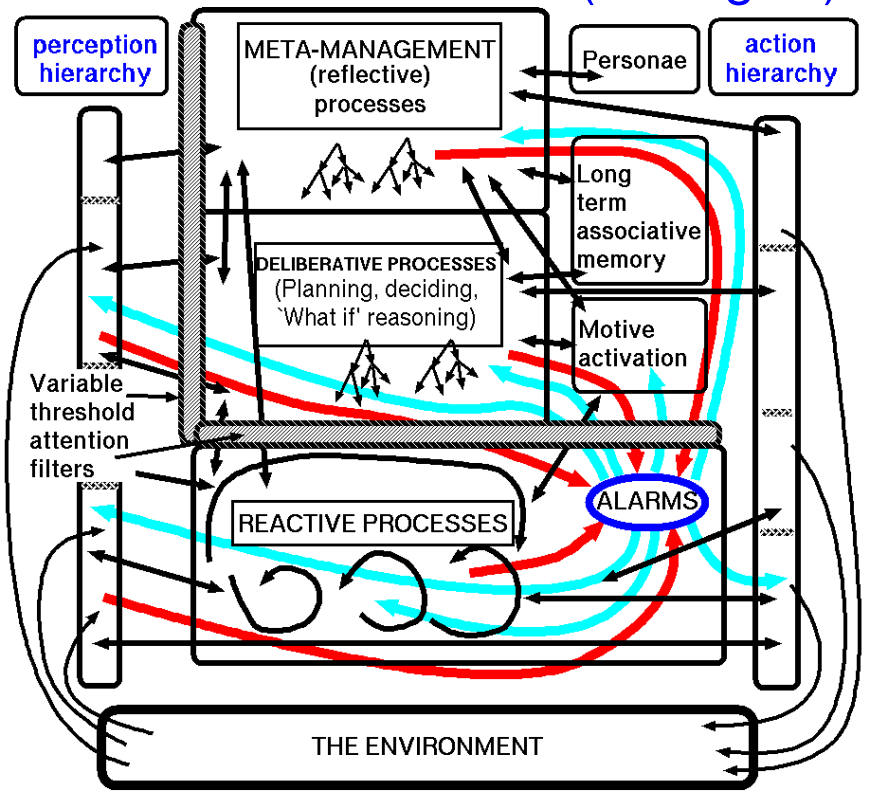
See also `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/whats-information.html`

We need to talk about architectures

Architecture Schema (CogAff)

Perception	Central Processing	Action
	Meta-management (reflective processes) (newest)	
	Deliberative reasoning ("what if" mechanisms) (older)	
	Reactive mechanisms (oldest)	

Architecture Instance (H-CogAff)



Many different sorts of architecture are possible within the CogAff schema, depending on what is in the various boxes and how they are connected.

H-CogAff is a special case: inspired by aspects of adult humans.

There are many other special cases, including microbes, insects, etc.

In humans, and some other species, the architecture grows itself after birth.

Varieties of individual development trajectories:
Cognitive epigenesis

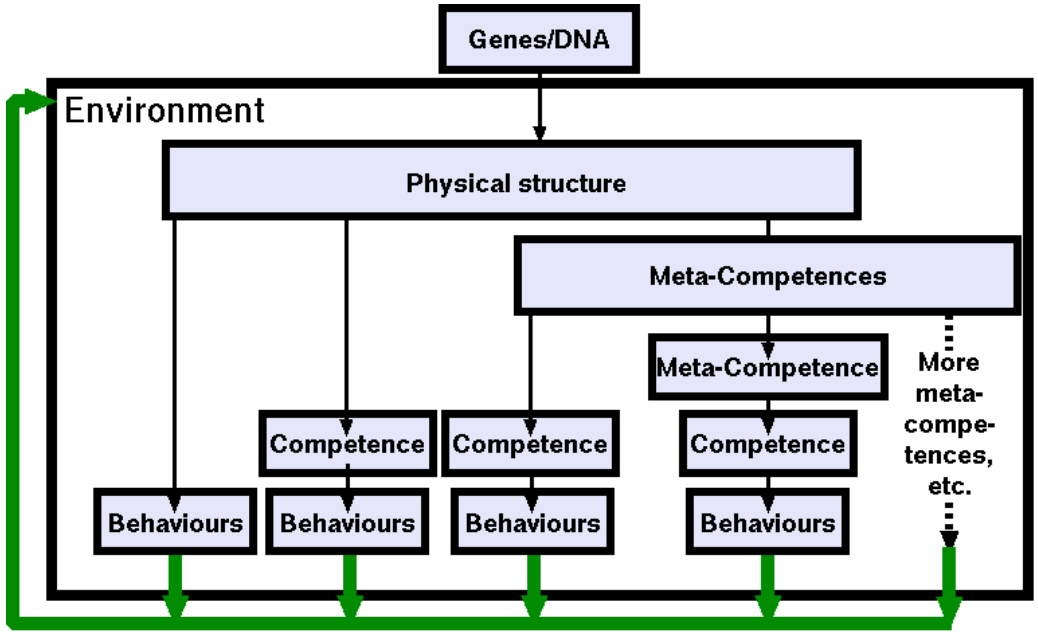
Multiple routes from genome to behaviours.

Individual trajectories (i-trajectories) involve various combinations of learning and development, based on various combinations of genetically provided competences,

including competences that provide new competences, at different levels of abstraction (meta-competences).

The more to the left a development process is the more it is genetically determined.

This is just a summary – explained in more detail in Chappell and Sloman (IJUC, 2007), Sloman and Chappell (BBS, 2007) and other online papers:



NB: All boundaries are somewhat fuzzy.
Environment affects all embedded processes.
For most species only the two leftmost routes are used.

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609`
Natural and artificial meta-configured altricial information-processing systems

Other papers also at at `http://www.cs.bham.ac.uk/research/projects/cosy/papers/`

Understanding architecture development and malfunction

- The human architecture has to grow itself while acquiring many kinds of orthogonal recombinable competences.
- We currently understand very little of what a normal human mind does and how it does it.

  For example something as common-place as seeing a 3-D environment in which everyday things happen is unexplained and we have no idea how to build robots that can do it.

  E.g. See `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cospal07`
- We cannot hope to understand how things go wrong if we don't know how they work normally:

  There are many ways such processes can go wrong and we currently understand only a tiny subset of them (e.g. how children can get confused about subtraction, or long division).
- Things can go wrong at all the different stages from conception to old age, only a tiny subset of which are now understood – most of which can be understood only within the design-based approach.

  How many psychoanalysts and psychotherapists will accept that?
  or the consequence that much treatment is actually a debugging process.
- When we've understood all the multiple sources of dysfunction : genetic, development, physical environmental, social, self-inflicted, etc. then we will be in a much better position to select appropriate treatments (debugging strategies) than we are now.
- In some cases treatment may be impossible.

Why don't brains repair themselves when damaged?

Every other part of the body can repair itself (to some extent) when damaged.

- When a bit of the brain is damaged (unless it's still very early) the information required to repair cannot be available in the (genetically determined) repair system.
- And it may be too late to re-acquire the information if it had to be acquired during a stage of development of other parts of the system.
- So developing techniques for inserting stem-cells and the like into damaged brains may be (mostly) doomed to fail.
- There could be some partial success, however.

Neuroscience has produced very impressive results – but

The great advances in neuroscience in the last few decades need to be treated with some caution:

- We (and neuroscientists) can be over-impressed because they have so much 'hard data' from detailed investigations of brains in many animals
- and they have acquired enormous confidence from impressive recent advances (e.g. using brain imaging)
- whose deep limitations have not yet become apparent to them!
- E.g. some of them proclaimed on the basis of such studies that there is a bifurcation of visual processing into two streams
  - ventral: concerned with what objects are in view
  - dorsal; concerned with where they are
- To a designer of working visual systems this is obviously untenable, since perception of large objects includes perception of their parts, including what they are and where they are.
- More recently (e.g. Milner and Goodale, 1992) it was acknowledged that the division was wrongly described: it's a division between
  - acquiring reusable enduring descriptive information about the scene vs
  - using transient visual information to to control current actions: visual servoing.
- However there are far more than two visual streams. see
  `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#pac07`

I am very grateful to Jaak Panksepp

For taking the trouble to write comments on the paper produced for this symposium. We have far more agreement than he suggests, and that is partly because what I wrote was not clear enough.

Some misunderstandings arose because I was only trying to address a small subset of prejudices and arguments about what is possible in AI (mostly related to the possibility of psychoanalysis), whereas he is trying to explain a vast cohort of phenomena that I did not mention (my paper was already over the requested limit) all of which I agree exist and believe need to be explained.

For example, we are in total agreement about the importance of chemical processes

> Molecular information processing is pervasive in biology, is the oldest kind, is all some organisms have, and is incredibly powerful: it is even used to build, and feed brains, and for feeding, monitoring and repair of most parts of the body.

We are also in agreement about the highly probable role of as yet undiscovered (types or instances) of highly complex strongly interacting dynamical systems.

> I have argued in another place that there are deep mysteries about the speed, diversity and creativity of human vision that probably requires entirely new kinds of dynamical systems, not the kinds Jaak talks about that are evolutionarily old but new kinds that are made up of many components created by the individual during a long learning process.

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#compmod07` (Continued on the last slide.)

Some (temporary) disagreements with Jaak

Given a few days (or hours) talking to him, I think I could persuade him that his notion of information-processing is too narrow, and that he is therefore not using a large enough collection of information processing concepts, especially multi-level control concepts.

I suspect he underestimates the variety of control mechanisms required for something like a human (or perhaps some other animals) to function, so whereas he suggests that I oversimplify by omitting things to be explained, I think he oversimplifies by underestimating the diversity of explanatory concepts and theories required. Interacting dynamical systems of the sorts he describes, insofar as they have a control function, can also constitute a virtual machine.

But other mechanisms are required for enjoying the study of philosophy or mathematics, or poetry.

Our disagreements probably arise because he focuses on mammals, and not enough on the diversity of problems solved by evolution (even tiny bugs). He appears to think cognitive mechanisms are more uniform and easier to implement than they they actually are: probably because he has not worked on trying to design a working visual system for a robot (e.g. a robot that can romp and play like young rats), or trying to model a mathematician capable of doing reasoning about geometry or about transfinite set theory.

I think he over-estimates the role of evolution in constructing the 'lower level' reactive sub-systems that operate in an adult animal: at least in the case of humans, and possibly some other animals, the fast automatic competences are products of development and training and to that extent created by the individual, and the environment, including the culture, e.g. musical sight-reading or athletic abilities.

He argues that in mammals there are basic emotions that are implemented in complex interacting fast changing dynamical systems.

But we need a theory of what sorts of dynamical systems are required for something to be described as having emotions and why they have to take just this form.

Otherwise we have only an empirical correlation between behaviours we interpret as emotional and certain neural dynamics, not an explanation of the connection. But I am convinced that if we talked for a while we would reach agreement because our basic approach is the same. And I would learn a great deal!

Related papers and presentations

The Computer Revolution in Philosophy

http://eprints.assc.caltech.edu/247/

Understanding causation in humans, animals and robots

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/wonac/

Do machines, natural or artificial, really need emotions?

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cafe04

What is information? Meaning? Semantic content?

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/whats-information.html

Four Concepts of Freewill: Two of them incoherent

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/four-kinds-freewill.html

The mind as a control system

http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#18

ASSC10 Poster: How an animal or robot with 3-D manipulation skills experiences the world

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0602

Orthogonal Recombinable Competences Acquired by Altricial Species

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0601

Two Notions Contrasted: 'Logical Geography' and 'Logical Topography'

Variations on a theme by Gilbert Ryle: The logical topography of 'Logical Geography'.

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html

More general collections

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/

http://www.cs.bham.ac.uk/research/projects/cosy/papers/

http://www.cs.bham.ac.uk/research/projects/cogaff/