Can we design a mind?

Aaron Sloman

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

Invited keynote talk at AID'02

Artificial Intelligence in Design conference,

Cambridge, July 2002

http://www.arch.usyd.edu.au/kcdc/conferences/aid02/

This presentation is available online as talk 15 at

http://www.cs.bham.ac.uk/~axs/misc/talks/

(Last changed July 20, 2002)

Acknowledgements

I am grateful for help from

Luc Beaudoin, Ron Chrisley, Catriona Kennedy,
Brian Logan, Matthias Scheutz, Ian Wright,

and other past and present members of the
Birmingham Cognition and Affect Group

and many great thinkers in other places

Related papers and slide presentations can be found at
http://www.cs.bham.ac.uk/research/cogaff/
http://www.cs.bham.ac.uk/˜axs/misc/talks/

This work is funded by a grant from the Leverhulme Trust for work on
Evolvable virtual information processing architectures for human-like minds.

ADVERTISEMENT
I use only reliable, portable, free software,
e.g. Linux, Latex, ps2pdf, gv, Acroread, Poplog, etc.
Diagrams are created using tgif, freely available from
http://bourbon.cs.umd.edu:8001/tgif/
I am especially grateful to the developers of Linux.

Abstract

Evolution, the great designer, has produced minds of many kinds, including minds of human infants, toddlers, teenagers, and minds of bonobos, squirrels, lambs, lions, termites and fleas. All these minds are information processing machines. They are virtual machines implemented in physical machines. Many of them are of wondrous complexity and sophistication. Some people argue that they are all inherently unintelligible: just a randomly generated, highly tangled mess of mechanisms that happen to work, i.e. they keep the genes going from generation to generation.

I'll attempt to sketch and defend an alternative view: namely that there is a space of possible designs for minds, with an intelligible structure, and features of this space constrained what evolution could produce. The CogAff architecture schema gives a first approximation to the structure of that space of possible (evolvable) agent architectures. H-CogAff is a special case that (to a first approximation) seems to explain many human capabilities.

By understanding the structure of that space, and the trade-offs between different options within it, we can begin to understand some of the more complex biological minds by seeing how they fit into that space.

Doing this properly for any type of organism (e.g. humans) requires understanding the *affordances* that the environment presents to those organisms – a difficult task, since in part understanding the affordances requires us to understand the organism at the design level, e.g. understanding its perceptual capabilities.

This investigation of alternative sets of requirements and the space of possible designs should also enable us to understand the possibilities for artificial minds of various kinds, also fitting into that space of designs. And we may even be able to design and build some simple types in the near future, even if human-like systems are a long way off.

Understanding complexity

Early AI theorists were over-optimistic about the likely rate of progress in AI, especially progress in emulating human capabilities, e.g. in vision, planning, problem solving, mathematical reasoning, linguistic communication, etc.

They grossly under-estimated the difficulty of the task.

Many critics of AI make the opposite mistake: claiming that goals of AI are unachievable.

Perhaps they over-estimate the difficulty.

The main problem is NOT shortage of computer power or limitations of computers.

The problem is that we do not know what the task is: we do not know what capabilities humans (and other animals) actually have.

The main problem is to know what the task is

- Merely saying that we want to build machines with human-like (or animal-like) capabilities assumes that we know what those capabilities are – whereas we don't – at least not yet, although we are learning, partly through doing AI and finding how un-human-like our systems turn out to be!
- Making progress requires a meta-level theory of what we need to know in order to specify those capabilities, so that we can then try to design systems that have them.
- We'll show that in part this requires us to find the right way to describe *the environment*.
- This leads to a circular bootstrapping process, in which doing AI helps us understand what the task is, by analysing the inadequacies of our early designs which surprise us.
- In addition we need a way to survey the space of possible designs for intelligent agents, so that we can understand alternative options available and see how humans are related to other organisms and machines.

What is it to understand how something works?

Often, understanding how a complex object works involves acquiring the kind of knowledge that a designer of the object has.

Example:

Understanding how a clock works involves knowing about

- **the source of energy,**
- **the mechanisms for transferring that energy to a time-indicating device,**
- **the mechanisms for regulating the flow in such a way as to produce the desired time indication.**

In general, a designer needs to understand a functional architecture.

When the object is an information-processing system, the task is more subtle because specifying the environment then depends in part on what information the object can process.

NOTE:
At the conference I was asked what I mean by "information" and "information-processing"?
The full answer is quite complex. Partial answers can be found in talk 4 and talk 6 here:
http://www.cs.bham.ac.uk/ axs/misc/talks/
Roughly: when you know the forms that information can take, the variety of contents it can have, the various ways it can be acquired, manipulated, analysed, interpreted, stored, transmitted, tested, and, above all, used, then you know (to a first approximation) what information is.
That knowledge grows over time, like our knowledge of what energy is.

Understanding an information-processing system

A designer of a working information-processing system, or someone trying to understand such a system, requires knowledge about the following:

– what the parts of the system are, and possibly how they are designed

  Understanding of a system may go down to a certain level, which is taken for granted.

  Some of the parts will contain symbols or other structures that express various kinds of information for the system. For instance, some parts may have information about other parts, as in an operating system. Some will have information about the environment.

– the relationships between the parts, including structural, causal, semantic, and functional relationships

  Functional relations are (roughly) causal relationships that contribute to some need, goal, or purpose, e.g. preserving the system.

– the subsystem of the environment with which the system interacts, and the structural, causal, semantic, and functional relations between the system and its environment.

These are all aspects of the architecture of the system: some are intrinsic aspects, while others are extrinsic.

These aspects need to be understood both by designers of systems and by scientists studying such systems.

Physical and virtual components, relations etc.

When we talk about components, inputs, outputs, causal interactions, etc. we are referring to phenomena that exist at various levels of abstraction, including components of virtual machines.

- **The components that we are interested in are not just physical components.**

  (They may include parsers, compilers, tables, graphs, schedulers, image interpreters...)

- **The various kinds of relations, properties, dynamical laws are not restricted to those investigated in the physical sciences**

  (not just physics, chemistry, astronomy, geology,... also relations like *referring to*, *monitoring*,)

- **We have to understand virtual machines at various levels of abstraction. This includes understanding how virtual machines interact with the physical world.**

  For example, when a chess playing program runs on a computer, the chess virtual machine includes entities and relationships like: kings, queens, pawns, rows, columns, colours, threats, moves of a piece, etc.
  These are not things that a physicist or chemist or electronic engineer can observe by opening up the machine and measuring things.
  Software engineers design, implement and debug virtual machines.
  Many people use virtual machines without realising that they do.

NOTE: action-selection in a *virtual* machine can cause changes in *physical* parts.

Ontological levels are everywhere

At all levels there are objects, properties, relations, structures, mechanisms, states, events, processes and CAUSAL INTERACTIONS.

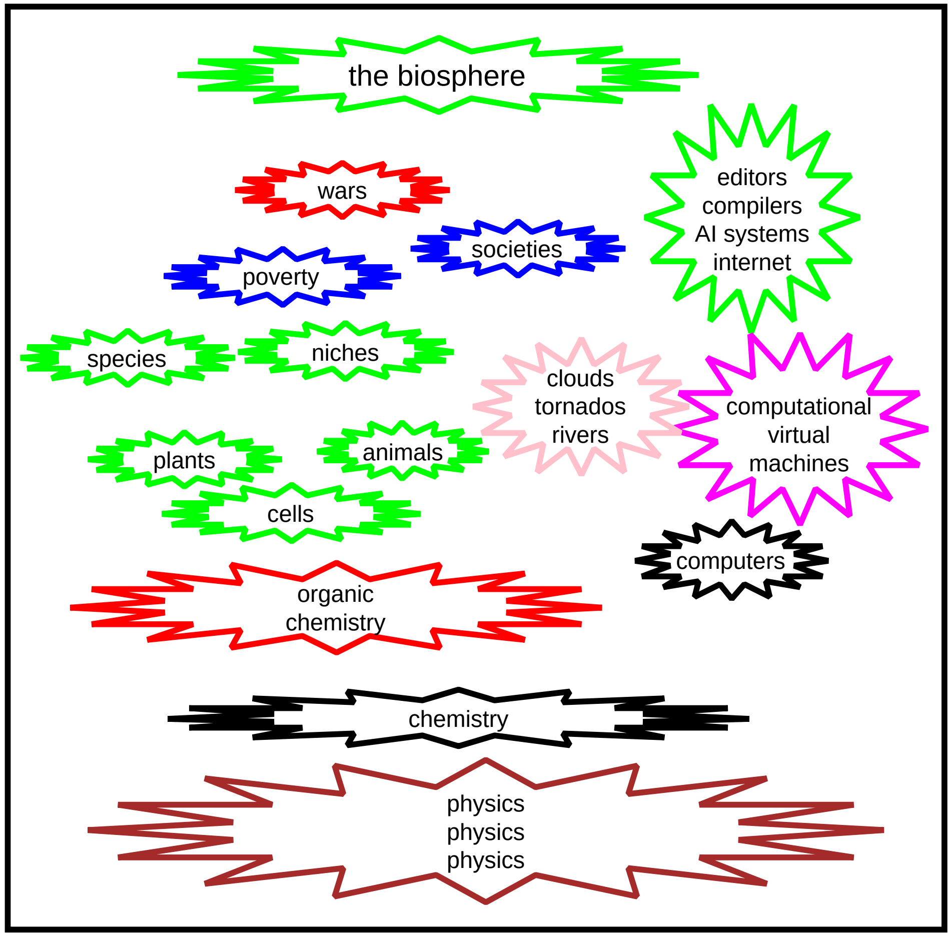
E.g. poverty can cause crime.

But they are all ultimately realised (implemented) in physical systems.

Different disciplines use different approaches (not always good ones).

Nobody knows how many levels of virtual machines physicists will eventually discover. (uncover?)

Our emphasis on virtual machines is just a special case of the general need to describe and explain virtual machines in our world.



See our IJCAI’01 Philosophy of AI tutorial for more on levels and causation:
http://www.cs.bham.ac.uk/˜axs/ijcai01/

Designing and understanding

Designing something which does not yet exist involves producing a specification for it, which can, in principle, be used as a basis for creating it.

Only "in principle" because sometimes a design presupposes some lower level mechanism whose design is not specified. If the set of such presuppositions is clear, then the design may be described as "relativised" to those presuppositions.
Most software designers do not know how to design the computers, compilers, or operating systems they take for granted. But generally they know that it can be done!

Understanding a working system that already exists involves acquiring the knowledge that *might have been used* in designing it.
(Usually making use of some presuppositions regarding lower level mechanisms.)

But that presupposes that the object has a design.

WHAT DOES THAT MEAN?

It does not mean that the object has a designer!

How can something have a design?

For an object O in environment ENV to have a design is for it to be characterisable using a description (which we sometimes refer to as a 'design', sometimes as an 'architecture specification', or just an 'architecture') specifying the following:

- **SC :** a set of enduring components (parts) of O
  (The contents of the set may change over time: parts may be created, destroyed or modified. Some may exist for only a very short time, others for longer.)
- **SR :** a set of relationships (including properties), applicable to members of SC
  (including structural, causal, functional, semantic relationships, some of which change over time)
- **ENV :** an environment, possibly itself of arbitrary complexity, including many entities, one of which may be O .
  (Typically ENV will itself have an architecture, as will many of its components. Some components of O may treat O , and components of O , as part of ENV .)
- **IP :** a set of types of inputs of various kinds from ENV to O
- **OP :** a set of types of outputs of various kinds from O to ENV
- **SLD :** a set of laws of dynamics for the system.
  (These may take many forms, including equations, rules and programs)

This is just a high level approximation. More details come later, based on O being an *information-processing* system.

Studying the environment is very important, and often very hard

It may be hard to find out what the relevant environment of an organism is without understanding the organism.

- If O has rich interactions with ENV (e.g. perceiving, acting, learning, communication) then understanding O requires understanding ENV .
- A special case: in order to understand a biological organism, one has to understand its *niche*.
- But what that niche is may be far from obvious, as we'll see.
- In particular, besides physical properties, a niche, or environment, for an information processing system may include what Gibson called "affordances" for that system. (J.J. Gibson, *The Ecological Approach to Visual Perception*, Erlbaum, 1986)
- The same physical environment may have different affordances for different organisms, or robots.
- E.g., representing the environment as a vector of measurements may fail to address the features of ENV that O perceives and uses.

So if O is an information-processing system, then understanding ENV requires understanding O and *vice versa*.

BOOTSTRAPPING IS REQUIRED!

Living organisms as information processors

In the case of most physical objects (e.g. a marble rolling down a helter-skelter) there is no clear separation between forces acting on it and information used by it.

The energy producing its motion comes from the forces determining what the motion (gravity, friction, collision forces, etc.) should be.

Moreover, the effects are generally direct and instantaneous.

(Exceptions are cases involving long-distance transmission and things like volcanoes that erupt, or dams that collapse, only after build-up of pressure, etc.)

Living things separate the acquisition of energy from the acquisition of information.

In both cases there can be significant delays between acquisition and use.

- Consumption of food provides energy for use later.
- Much information is stored for use later.

In the simplest living things information from the environment is not stored, but used immediately. Nevertheless it does not always directly drive the "motors". Rather information is used to switch on and direct internal energy supplies.

Information and energy in living organisms (intelligent systems)

Typically, in such systems:

- Sensors obtain information, on the basis of which (together with previously stored information) actions are selected. (Both external and internal actions).

- Previously stored energy (mainly chemical energy) provides the forces required to perform the actions.

One type of evolutionary development makes the information processing more and more complex, flexible, varied, powerful, so that more and more energy is required to support internal information processing actions as well as external actions.

The result is a type of organism that has to be high up a food pyramid, and is expensive to produce and costly (for the gene pool) to lose.

There cannot be many such species: the vast majority of species have numerous, low cost, relatively unintelligent, individuals.

More on the Inputs and Outputs

We can break down the specifications of inputs and outputs of the object or organism O in more detail.

IP : the set of inputs may include various subsets:

- **SEN :** sensors acquiring information of various types from ENV
- **ING :** ingestors acquiring matter from ENV
- **EPORTS :** ports through which energy can be acquired.

Sometimes these overlap: a mother pulling a child by the hand is providing both information about the direction to move in and some of the energy propelling the child! Ingesting food is also ingesting matter, partly to provide energy, partly body-parts.

OP : the set of outputs, may include

- **SIG :** a set of signallers transmitting information from O to ENV
- **MAT :** a set of matter output mechanisms transmitting (exuding) matter, including reproductive output, and waste possibly, from O to ENV , or from internal storage to components of O .
- **MOT :** a set of motors transmitting energy from O to ENV (e.g. wings, legs, jaws, wheels, hands, claws, ...)

As in the mother/child example, these can overlap.

Notes on inputs and outputs

- The same physical component can function both as part of the motor system and part of the sensor system, e.g. hands, tongue, etc.
- In some systems the inputs and outputs (and internal state changes) all vary continuously.
- In some they are all discrete.
- In some they are a mixture of discrete and continuous (digital and analog mechanisms.)
- Sometimes a particular sensor takes in information of many types, processed at different levels of abstraction.

  E.g. in humans, eyes feed in information of many kinds encoded in patterns of photons, including: information about physical structures, causal and functional relations (affordances) the states of mind of other agents, written information, etc. The same physical transducer is shared between many types of perceptual information processing subsystems.

  See A. Sloman (1993) The mind as a control system, in *Philosophy and the Cognitive Sciences*, Eds. C. Hookway & D. Peterson, Cambridge University Press, pp. 69–110.
  Also online at http://www.cs.bham.ac.uk/research/cogaff/

- There is no requirement that all of the processing within O should be identifiable by observing input/output relations.

  E.g. there may not be sufficient output bandwidth. Moreover, significantly different internal processing can produce the same input/output mappings. So testing theories about what is happening within O (within the virtual machine architecture) typically requires far more than experimental observation of O's behaviour.

Understanding the environment

One of the hardest tasks in attempting to design a mind for an object O is to specify the environment ENV correctly.

- It may appear that that is trivial: the environment is just the physical world in which O is embedded.
- However that is clearly misleading when O is an intelligent system and ENV includes communications from other agents. In that case, perceiving the environment involves understanding the communications.
- More generally, as Gibson pointed out, environments include "affordances" of various kinds.

  A physical object such as a table, in addition to having geometrical and physical properties, may provide positive and negative affordances, such as
  - support (for a person leaning on the table, or a cup placed on the table by a person)
  - obstruction (for a person wishing to cross the room).
- Which affordances ENV has when perceived by O will depend on
  - The structure of ENV and its laws of behaviour
  - What O can desire or need
  - What O can do, i.e. the types of actions it can perform (e.g. grasping, pushing, jumping)
  - O 's information processing capabilities, e.g. its perceptual capabilities.
- The affordances for different organisms are different: they can inhabit different niches even when they are at the same location.

Finding the right ontology for ENV

Without understanding the needs, action capabilities, and information processing capabilities of an organism, we cannot describe its environment in a manner that is relevant to understanding its information processing architecture: its mind.

This involves developing an appropriate ontology for ENV .

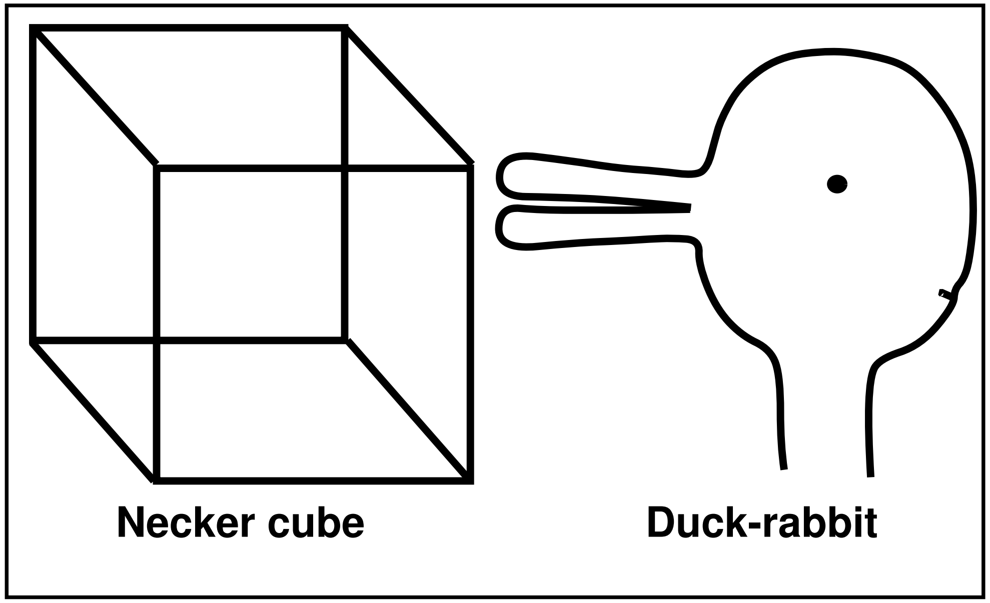
Likewise if we wish to design a synthetic agent we need to be clear about its affordances, and the resulting ontology of its environment. Otherwise the requirements for the design will be under-specified.

- Finding the right ontology (including the correct set of affordances) in both animals and robots may be very difficult: scientists and engineers may be ontologically blind

  (See Sloman and Chrisley's WGW02 paper at http://www.cs.bham.ac.uk/research/cogaff/)
- Much of the history of AI has involved researchers making over-simple assumptions about the nature of the environment, and the nature of the perceptual and action processes required by an intelligent system.
- It is often more appropriate to describe the environment in terms of its syntax, and perception as involving parsing and interpretation, than to think of the environment as having physical properties and perception as measurement.

Levels in perceptual mechanisms

Seeing the switching Necker cube requires a grasp of geometrical properties and relations, as well as connectivity relations.
Seeing the flipping duck-rabbit uses far more subtle and abstract percepts, going beyond geometric and physical properties.
(Compare Marr on vision)



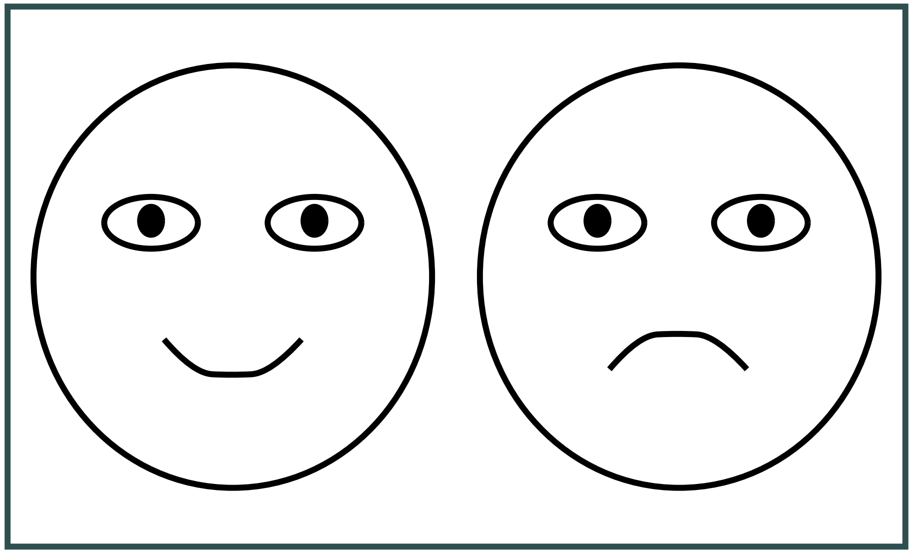
Things we can see besides geometrical properties:

- Which parts are ears, eyes, mouth, bill, etc.
- Which way something is facing
  (What does that mean? Why might it be important for prey or for predators?)
- Whether someone is happy, sad, angry, etc.
- Whether a painting is in the style of Picasso...
- Whether something is graspable, and if so how
- Which subsequent movements are facilitated by different ways of grasping it.

Seeing Faces

Seeing facial expression as we do may just be a very old and simple process in which features of the face trigger reactions in a pattern-recognition device.

Or it may also involve deployment of sophisticated concepts that developed only through the evolution of meta-management.
(Explained later)

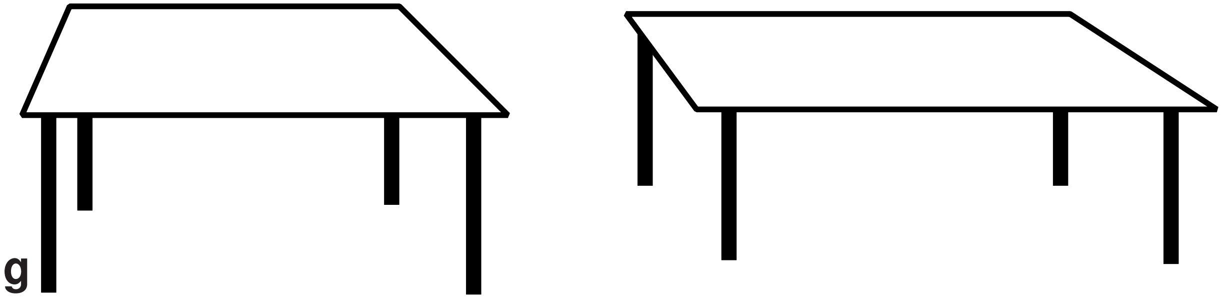


Some people see one pair of eyes as “looking happy” while the other pair “looks sad” or “looks angry”. (A context effect.)

For more on levels in perceptual mechanisms see talks on vision and visual reasoning here: http://www.cs.bham.ac.uk/˜axs/misc/talks/

Seeing tables

What sorts of affordances does a table provide?



– Obstruction
– Support
– Pulling, lifting, pushing, in various ways depending where you hold it and how.
– Easy availability of a collection of tools or papers, etc., in easy reach
– Social cohesion during meals
– Types of construction and repair methods
– ....

(See my ‘Actual possibilities’ paper at the CogAff web site.)

Some of the affordances are conditional: e.g. you can pull the table if you (a) move closer and (b) grasp a leg or the edge.

How do we (and other animals) represent collections of possibilities and constraints on possibilities? How do we use our grasp of such possibilities and constraints to work out what to do?

Do we, or chimps, use modal logics?

More on the components of O

SC , the (changing) set of enduring components (parts) of O , will need to include:

- **Short term and long term information stores of various kinds, including a variety of stores for temporary perceptual information at various levels of abstraction (e.g. physical details and affordances). Longer term stores will include both information about specific objects and events, and also re-usable generalisations.**

  Information may be encoded in physical or in virtual machine structures, which may vary continuously or discretely.

- **Mechanisms for analysing, interpreting, manipulating and deriving information, by operating on the structures that encode the information.**
- **Mechanisms for generating or activating "springs of action": i.e. goals, desires, preferences (motivators).**
- **Mechanisms for changing components in various ways, i.e. mechanisms for correcting mistakes, for learning and for development.**

The need for meta-information

If some part of O uses a type of formalism as an information bearer, then the specification of that formalism's syntax and semantics may be implicit in the mechanisms that operate on the formalism.

Example: the syntax and meaning of bit patterns representing pointers, numbers and machine instructions for a computer is typically implicit in the digital circuitry and firmware that operates on pointers and instructions.

Alternatively there may be explicit meta-information saying what the formalism is, what its syntax is, and how meanings of complex structures are determined by meanings of their simpler components.

Example: a compiler designer for a computer must have an explicit specification of the syntax and semantics of the machine code implemented in bit-patterns.

Different forms of representation with different kinds of syntax and semantics may be used in different parts of a complex system.

We don't really know how many different types of formalisms are used in human minds, nor which have only implicitly specified syntax and semantics and which have explicit specifications of how they can change over time.

Some of the formalisms will have to be only implicitly specified, to avoid an infinite regress of interpretations.

Ontologies within

Insofar as percepts, beliefs, desires, intentions, fears, refer to entities in the environment ENV , or entities within O , O will have to use one or more ontologies.

This ontology may be implicit in the mechanisms and processes, or explicitly specified.

The ontologies presupposed by the lowest level mechanisms must be implicit in the mechanisms.

E.g. a computer implicitly uses an ontology including registers, memory locations, addresses, instructions, one or more instruction pointers, numbers, arithmetic operations, input and output devices, etc.

Typically different parts of the system will use different ontologies. E.g. your posture control sub-system will not need an ontology referring to days of the week.

The ontology, or ontologies, used by O can change over time. This is a significant aspect of normal human development. The extent and diversity of ontology construction may be a uniquely human feature.

New-born infants show no sign of using an explicit ontology for the environment, yet typical eight year olds have, at least within some portion of the architecture, a rich collection of reportable explicit information about what sorts of things exist in the environment and in the child, e.g. animals, clouds, dreams, pains, etc.

More on the relationships between components

SR : a set of relationships (including properties), applicable to members of **SC** (including structural, causal, functional, semantic relationships)

.... to be extended ....

More on the laws of dynamics

SLD : a set of laws of dynamics for the system, specifying

- how the membership of the above sets changes over time
- how the various types of components, inputs and outputs behave
- including how percepts are formed, beliefs are created or modified, new conclusions are derived, motives are generated, and so on
- how properties of and relations between components change, e.g. development of new internal communication channels.

Laws of dynamics may be specifiable at different levels of abstraction, including

- laws of physics and chemistry
- laws of physiology or digital circuitry
- laws of various virtual machines in a tower of implementation levels.

Example:

- An information processing system may process different kinds of information at different times, and may process information in different ways (e.g. looking at brush-strokes in a painting as opposed to looking at the scene depicted).
- These changes involve changes of attention
- The mechanisms that switch attention will typically be components of virtual machines, involving goals, desires, preferences, reasoning capabilities, etc.
  Most of the laws of behaviour have yet to be discovered and formulated.

Varieties of dynamical laws

- Laws specifying classes of internal or external behaviours can take many syntactic forms including equations, various kinds of computer programs, sets of condition-action rules, sets of constraints and rules for constraint propagation. Some of them specify only serial behaviours (e.g. conventional computer programs and condition-action rules), while others specify concurrent processes, e.g. event-driven programs, operating system specifications.
- Some laws are implicit in mechanisms. Others may be explicitly formulated within O , and therefore more directly subject to inspection and change by O .
- Some of these require rich semantic contents, for instance condition-action rules referring to complex conditions and actions.
- Some of the laws may specify how other laws change. For instance, during sleep, or drug-induced states, some normal laws are temporarily replaced by others. During learning and development permanent changes may occur, though very little is known about this, e.g. how a child acquires the ability to think about infinite sets.
- If O 's goals or needs, or information processing capabilities change, then the relevant affordances in environment ENV will also change. I.e. internal changes in virtual machines can induce external changes.
- Insofar as O has components of very different kinds, e.g. both reactive and deliberative components, the laws and what they refer to may be very different in form and content.

Emergent laws

Laws that specify dynamics of virtual machine components may not be derivable from or verifiable from external observation of input-output mappings.

They may also not be derivable from underlying physical/chemical laws.

Even when virtual machine VM is implemented in physical machine PM it may be impossible to derive the laws of VM from those of PM using *only* logical and mathematical principles.

- **The concepts used to describe VM and formulate its laws, may not be definable in terms of the concepts relevant to PM alone.**
- **Therefore statements containing such VM concepts will not be derivable from statements using only concepts required to describe PM.**

For more on emergent ontologies, and a discussion of circular causation between levels of abstraction, see the Sloman-Scheutz IJCAI'02 tutorial on philosophical foundations of AI
http://www.cs.bham.ac.uk/~axs/ijcai01

Motive dynamics

Generation and processing of motives requires a variety of mechanisms.

- There are many sorts of motive generators: MG
  - Some triggered by internal states, e.g. temperature, fluid requirements, mating mechanisms
  - Some triggering implicit motives, others explicit motives.
  - Some triggered by percepts, e.g. seeing danger or an opportunity
  - Some (in humans) triggered by thought processes, e.g. remembering something
- However, motives may be in conflict, so motive comparators are needed: MC.
- But over time new instances of both may be required, as individuals learn, and become more sophisticated:
  - Motive generator generators: MGG
  - Motive comparator generators: MCG
  - Motive generator comparators: MGC
  - And maybe more:
    MGGG, MGGC, MCGG, MCGC, MGCG, MGCC, etc ?

A more complete analysis would need to distinguish implicit and explicit motives, generators and comparators. An explicit version includes some structure within the system that can be inspected, modified, stored, compared with others, etc.

There are also evaluators:

Evaluators are required in order to determine whether some current or future possible state should be preserved, enhanced, reduced, terminated, sought in future, etc.

- Current state can be evaluated as good, or bad, to be preserved or terminated (or intensified or reduced).
- Evaluations may interact with learning in some architectures (e.g. positive and negative reinforcement).
- These evaluations can occur at different levels in the system, and in different subsystems.
- This can account for many different kinds of pleasures and pains.
- “Error signals” form a special case

NOTE: Evaluations are often confused with emotions.

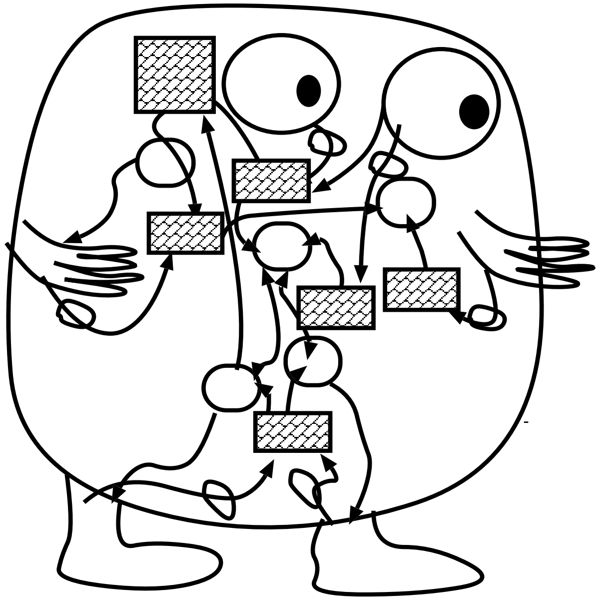
For more on this see talks on emotions and affect in this directory

http://www.cs.bham.ac.uk/˜axs/misc/talks/

and papers in this directory

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

Could the architecture be an unintelligible mess?



Some people argue that we cannot hope to understand products of millions of years of evolution. They work, but do not necessarily have a modular structure or functional decomposition that we can hope to understand.

YES, IN PRINCIPLE.

BUT: it can be argued that evolution could not have produced a totally non-modular yet highly functional brain.

Problem 1: time required and variety of contexts required for a suitably general design to evolve.

Problem 2: storage space required to encode all possibly relevant behaviours if there's no "run-time synthesis" module.

Conjecture: evolution, like good engineers, 'discovered' the virtue of re-usable modules and and nearly decomposable complexes (H.A.Simon 1967).

Why ‘architecture?

Once upon a time, insofar as AI studied mechanisms they were mainly thought to be

– representations

and

– algorithms.

(Or that’s what people thought they thought – so they wrote it in textbooks. Of course, knowledge had to be added, using the representations – logic, lists, trees, graphs, arrays, ...)

More recently (since mid/late 1980s?) it has become clear(er) that we also need to understand ways of putting things together, possibly in large and complex systems, often with many things going on at once.

So we need to study architectures

Why study “architectures” (plural) ?

Even for someone whose primary motivation is to understand human minds, it is necessary to investigate diverse architectures.

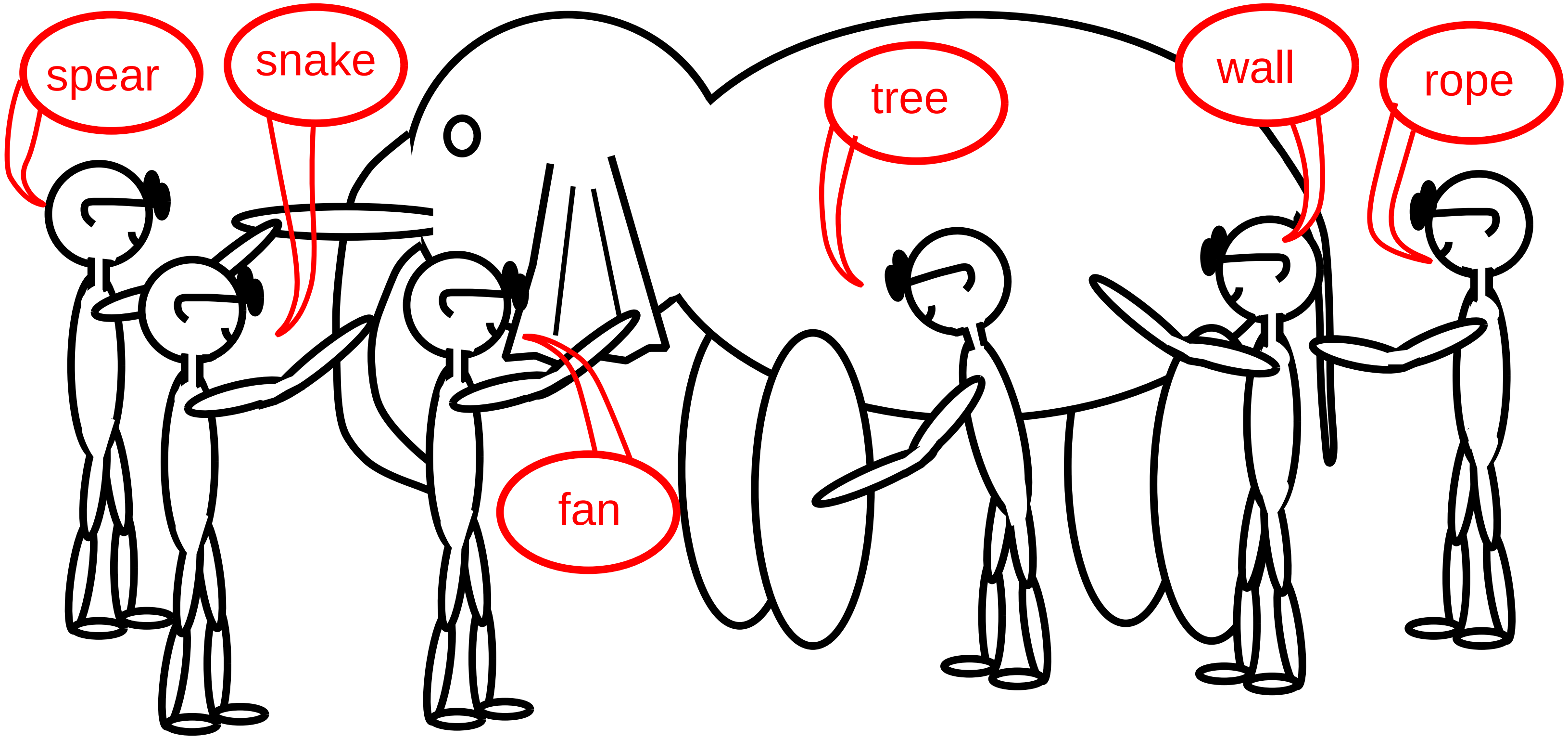
- Because there is not one human architecture, but many (infants, children, various kinds of people with brain damage).
- Because one aspect of individual human learning and development from infancy is “bootstrapping” a succession of new architectures from old ones, e.g.:
  - adding new components
  - adding new links
  - adding new forms of representation and inter-component communication
- Because our architecture is a product of co-evolution with many other co-evolving architectures helping to shape it (including our ancestors, who have left bits of themselves in us).
- Above all because you don’t understand one thing until you compare it with others, investigate the similarities and differences, and analyse their implications

  i.e. we need to understand trade-offs in a design in order to understand the design.

WE SHOULD AT LEAST TRY TO SEE THE WHOLE ELEPHANT

What is an Elephant?

See: “The Parable of the Blind Men and the Elephant”
by John Godfrey Saxe (1816-1887)
http://www.wvu.edu/~lawfac/jelkins/lp-2001/saxe.html



Who can see the whole reality?

...continued

We can hope to see "the whole elephant" more clearly if we understand the variety of processes that can occur within a human information processing architecture.

Moreover, most mental concepts are (I claim) architecture-based and 'polymorphic', so

by looking at different architectures, for human adults, for children, for dogs, for rats, for fleas....
we may understand the even larger variety of affective states
and processes that different architectures support

and thereby get a clear grasp of possible meanings for words like "emotion" and other mental words.

There are many "elephants" for us to study.

Many other familiar mental concepts are polymorphic cluster concepts, e.g.

"CONSCIOUSNESS", "BELIEF", "INTENTION", "INTELLIGENCE", "PLEASURE", "PAIN", "FREEDOM", ETC.

and can be refined and clarified in an architectural framework.

Understanding alternatives

- Most of the the problems that are of interest in designing minds have different solutions depending on which sort of mind we are considering, in which sort of environment, or niche.
- One way to think about this is to think of the space of possible designs and the space of possible niches as linked by descriptions of ways in which different designs match a particular niche and ways in which the same design matches different niches.
- This gives more information than the use of fitness functions for designs that produce a number, or an ordering of designs.
- Since mismatches can produce pressures for changes in designs, and this can produce new niches, leading to new kinds of matches and mismatches, we have interacting systems concurrently tracing trajectories through design space and through niche space with complex interacting feedback loops. A more mathematical formulation of this is desirable. However it is possible that a new type of mathematics is required.
- The next two slides illustrate these ideas graphically.

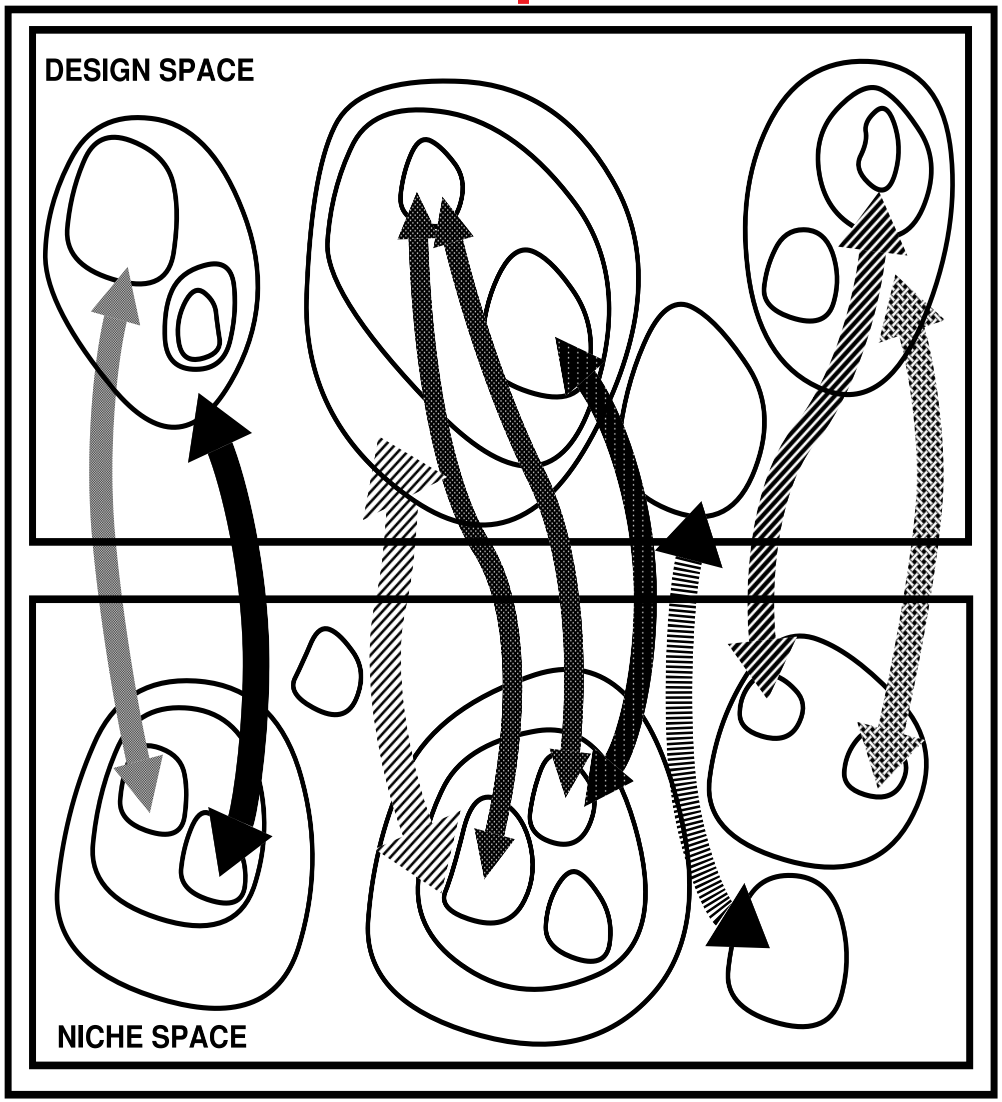
Design space and niche space

There are discontinuities in both design space and niche space: not all changes are continuous (smooth).

Many researchers look for one “big” discontinuity (e.g. between non-conscious and conscious animals).

Instead we should investigate many small discontinuities as features are added or removed.

A continuum (smooth variation) is not the only alternative to a big dichotomy.



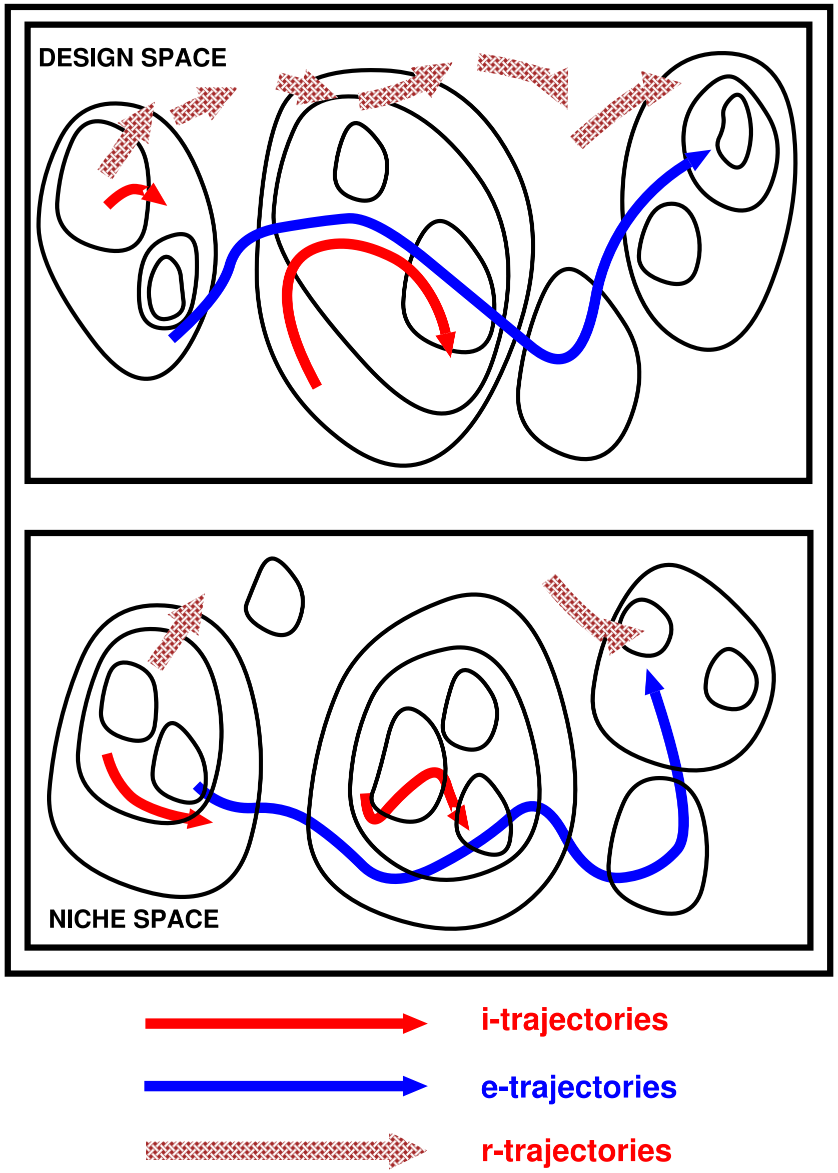
The arrows represent fitness relationships between designs and niches. The relations different in kind: they are not simply numerical values, but can include structural descriptions, vectors of values, etc..

Trajectories in design space and niche space

There are different sorts of trajectories in both spaces:

- **i-trajectories:**
  Individual learning and development
- **e-trajectories:**
  Evolutionary development, across generations, of a species.
- **r-trajectories:**
  Repair trajectories: an external agent replaces, repairs or adds some new feature. The process may temporarily disable the thing being repaired or modified. It may then jump to a new part of design space and niche space.
- **s-trajectories:**
  Trajectories of social systems.

Some e-trajectories may be influenced by cognitive processes (e.g. mate-selection). We can call them c-trajectories



Need for new ways of studying dynamics

- The type of evolutionary process described here includes feedback loops involving multiple discontinuous trajectories in at least two different spaces.
- Do we have the right conceptual tools to study the dynamics?
  - E.g. we need to clarify the nature of different kinds of states: can affective states and non-affective states be distinguished in a principled way.
  - What are cognitive states?
- Can we understand the full variety of ways in which information can be encoded (in physical and in virtual machines), manipulated and used.
- Will we need new mathematics? (The limits of equations, and programs).

(In part the answer will depend on whether we can even study small regions of design space and niche space fruitfully – as Matthias Scheutz has been doing.)

Demonstrations available:

- **Reactive systems**
  - **Flocking behaviour** Blindly following a “leader” by reacting only to sensory input
  - **Emotive reactive system** emotional states produced by different percepts alter behaviour
  - **The sheepdog demo** It has different global states with different collections of reactions
  - **Eliza** A reactive system where reactions include instantiated variables
- **A deliberative system**
  - **A blocks world conversationalist** Loosely modelled on Winograd’s SHRDLU (1971)

Exploring architectures and their implications teaches us to abandon simple classifications of systems, and simple classifications of the processes that can occur in them.

Is a ‘principled’ investigation possible?

It is possible that the spaces and trajectories are too messy to be investigated in any other way than to examine particular cases in great detail.

But perhaps there is a way of being more principled:

- **Investigate “dimensions” in which architectures (designs) can vary.**
- **Investigate “dimensions” in which niches, sets of requirements, problems, etc. can vary**
- **Investigate the variety of relationships between designs and niches:**
  - **e.g. is it all just numerical fitness functions?**
  - **What’s the alternative.**
- **Try to classify and model the different kinds of dynamics involved.**

Some of that may require development of new kinds of computers, or new non-computational mechanisms – so what?
Physicists have never tried to define their field by the formal tools available to them at a particular time.
Neither should we: start from problems not tools.

(Both change over time.)

Towards a unifying theory of architectures

- We need good general-purpose concepts for describing and comparing different classes of architectures for organisms and robots, and possibly other things.
- We build up our concepts by relating them to a space of possible architectures for integrated (non-distributed) agents.
- This space is characterised by a generic schema (a sort of grammar) specifying types of components and ways in which they may be related.

In the following slides we present a schema called CogAff. It is only a tentative first draft and will certainly have to be enriched.

We do this by

– presenting different perspectives for dividing up an architecture

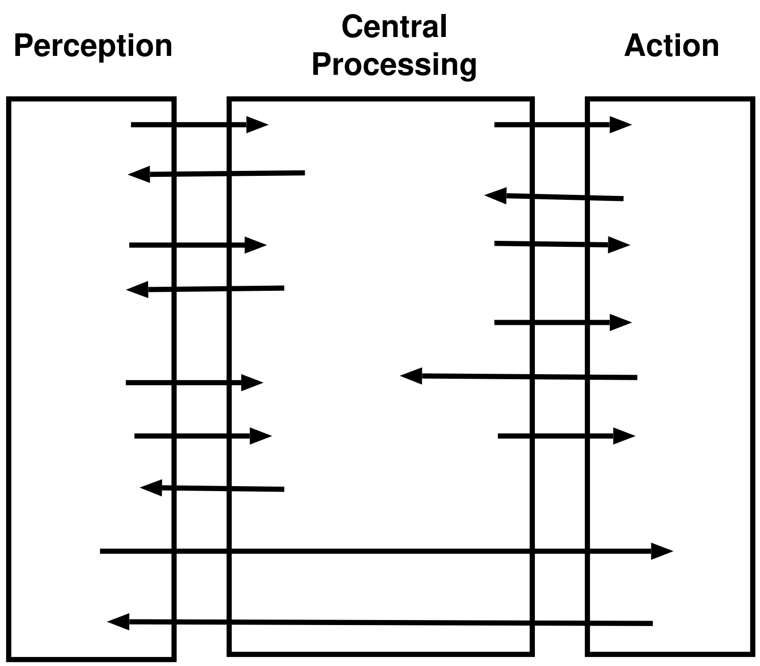
– showing how to overlay those perspectives to get a deeper understanding of the diversity

– indicating in a sketchy way how various aspects of human minds and other information processing systems relate to the various divisions in the architecture.

NOTE:
CogAff does not cover multi-agent architectures except insofar as the components of a single integrated architecture can be viewed as agents.

Perspectives on complete agents

1. The "triple tower" perspective



(Many variants – Nilsson, Albus, ...)

"Nearly decomposable" systems. (H.A.Simon)

Boundaries can change with learning and development.

The main basis for distinguishing central from perceptual and action mechanisms: causal influence.

- The contents of the perceptual tower are largely under control of input from sensory transducers. Their function is primarily to analyse and interpret incoming information. They may also be 'in registration' with collections of sensory transducers.
- Similar criteria can be used for specifying contents of action tower.
- Contents of 'central' tower (a) change on different time-scales from those of perceptual and motor towers (b) are not closely coordinated with them.

A less obvious perspective

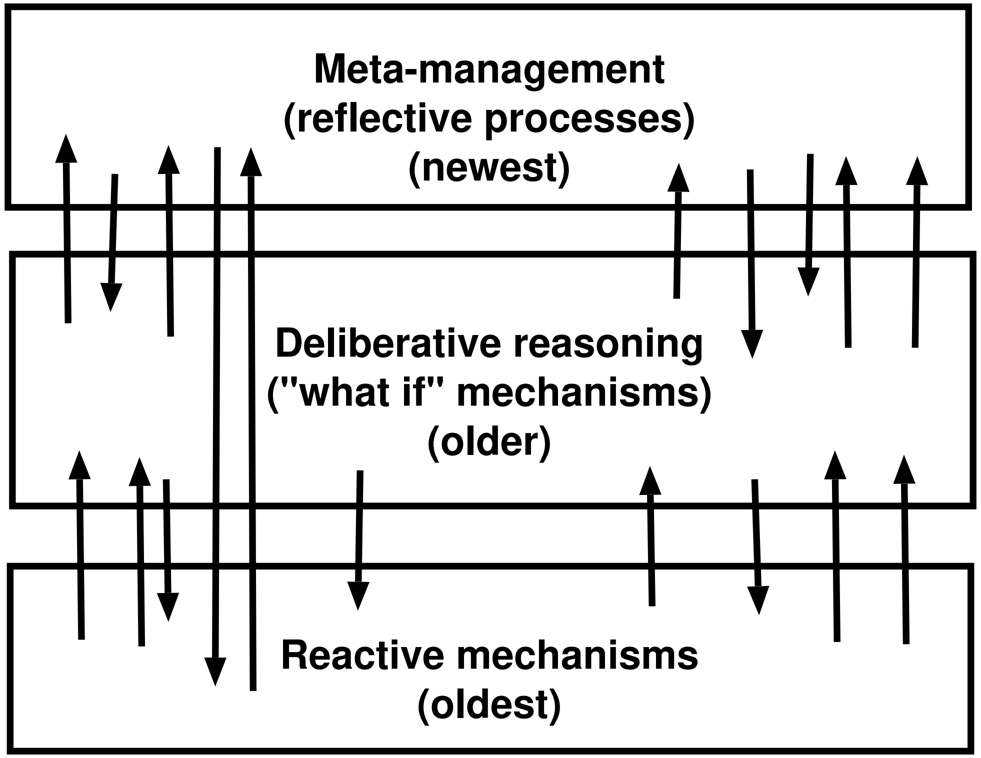
2. The "triple layer" perspective

Another common architectural partition (functional, evolutionary).

There are many variants – for each layer.

All mechanisms must be implemented at some level in reactive systems.

Some people separate reflexes from more complex reactive mechanisms which include internal state changes.



Reactive mechanisms

- They are very diverse and may include many concurrent sub-systems, including "alarm" mechanisms.
- They may include forms of learning or adaptation.
- Some reflexes (innate or learnt) connect sensors directly to motors.
- Reactive mechanisms can be highly parallel and very fast.
- They may use analog or digital components or both.
- They may include neural nets, condition-action rule systems, lookup tables, decision nets, and other mechanisms.
- Some reactions change only internal state, affecting future reactions.
- Some internal states may act as goals.
- It may be difficult or impossible to program them directly or provide explicit information for them to use (compare neural weights.)
- They make possible 'alarm-driven' primary emotions.

NOTE: In principle any form of externally observable behaviour over any time scale can be produced by a reactive system.

However, satisfying the same set of true counterfactual conditionals as a human deliberative system may require an impossibly large information store and an impossibly long and varied process of evolution and training.

Deliberative mechanisms

- Can represent and reason about non-existent or future possible entities.
- Some can also reason about what might have been the case in the past.
- They allow alternative options to be constructed, evaluated, and compared.
- They can vary in the representational forms they use and the sophistication of their semantics.
  - Simple deliberative mechanisms may use only one step lookahead, and very simple selection mechanisms.
  - More sophisticated versions use compositional semantics in an internal language whose grammar admits unbounded complexity.
- They require a re-usable general purpose working memory (garbage collectable?)
- They require stored generalisations about what actions are possible in particular situations, and about the consequences of actions.
- They may be able to learn (new formalisms, new ontologies, new associations, ...)
- They benefit from perceptual systems that produce high-level chunked descriptions of the environment
- They may be able to train reactive systems that cannot be directly modified.
- Typically slow, serial, resource limited. (Why?) May need attention filter.
- They make possible secondary emotions using global ‘alarm’ mechanisms linked to deliberative mechanisms.

Meta-management mechanisms

- They can monitor, categorise, evaluate, and (to some extent) control other internal processes – e.g. some deliberative processes, or some perceptual processes. (See Barkleys's 1997 book on ADHD)
- This includes control of attention, control of thought processes. (Control which is lost in tertiary emotions.)
- They can vary widely in sophistication, e.g. depending on social learning.
- They require concepts and formalisms suited to self-description, self-evaluation
- They support a form of internal perception which, like all perception, may be incomplete or inaccurate, though generally adequate for their functional role.
- The concepts and formalisms may be usable in characterising the mental states of others also.
- Different meta-management control regimes may be learnt for different contexts (different socially determined “personae”).
- Evolution of sensory qualia: occurs when it is useful for meta-management to look inside intermediate levels of perceptual processing (why?).
- If meta-management mechanisms are damaged, blind-sight phenomena may occur. (Experiments requiring subjects to *report* what they see typically use the meta-management layer! What’s happening in other layers may be unnoticed.)

Varieties of meta-management

- There may be different types of meta-management using more or less sophisticated forms of representation and processing.
- They can also vary in the types of evaluation they can apply
- In humans much self-categorisation and self-evaluation is socially/culturally determined. (E.g. feelings of guilt or sin)
- The existence of meta-management may provide a "niche" encouraging evolution of higher level *perceptual* mechanisms categorising mental states of other agents. (Top-left box in grid diagram. Likewise top-right box for action mechanisms.)
- This may have required parallel evolution of involuntary "expressive" behaviours (Sloman 1992 on the dangers of complete voluntary control of sincerity.)
- The absence of meta-management was a major factor in the fragility and incompetence of many old AI systems (e.g. they could not tell when they were reasoning in circles, or solving a minor variant of a previously solved problem.)
- Mechanisms for triggering and modulating meta-management processes may produce a far wider variety of affective states than scientists have so far categorised. (Contrast novelists!)

More on the three layers

The layers differ in:

– **Evolutionary age** (reactive oldest).
– **Level of abstraction of processing** (reactive least abstract),
– **The types of control functions, and mechanisms used**
  (e.g. ability to search, evaluate, compare; amount of parallelism; use of neural vs "symbolic" mechanisms)
– **The forms of representation used**
  (e.g. flat vs hierarchical compositional syntax and semantics)

The distinctions between layers are not necessarily very sharp, and there can be intermediate cases.

In fact it is very likely that evolution produced intermediate cases.

Layered architectures have many variants

Later we'll see that designers present layered architectures with different subdivisions and different interpretations of subdivisions, and different patterns of control and information flow.

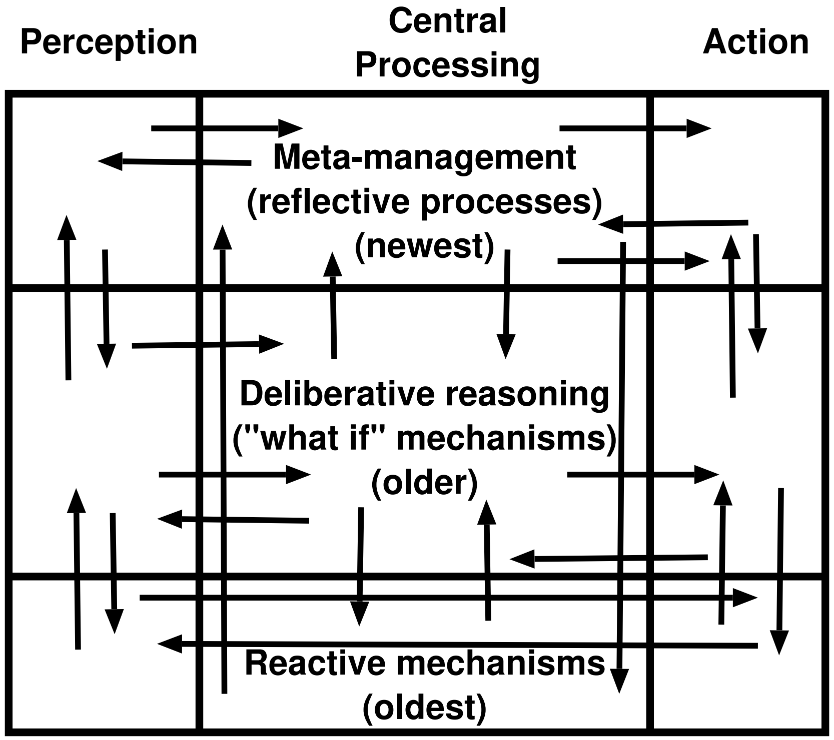
Divisions between layers can be based on:

- evolutionary stages
- levels of abstraction,
- control-hierarchy, (Top-down vs multi-directional control).
- information flow
  (e.g. the popular 'Omega' Ω model of information flow, described below)

We'll try to present CogAff as subsuming many such design options.

LAYERS + PILLARS = CogAff GRID

We can overlay the two views, giving a grid of co-evolved sub-organisms, each contributing to the niches of the others.



THIS IS AN ARCHITECTURAL "SCHEMA" SPECIFYING POSSIBLE COMPONENTS AND RELATIONSHIPS BETWEEN COMPONENTS, NOT AN ARCHITECTURE.

The CogAff schema defines a variety of components and linkages.
Not all the components, and not all the communication links, need be present in all species of natural or artificial architecture.

It does NOT specify control flow, or dominance of control: many options left open.

Information may flow in ways not shown by the arrows - e.g. diagonally across layer boundaries. (Example?)

This is a very general schema.
Contrast the H-Cogaff (human) instance (below).

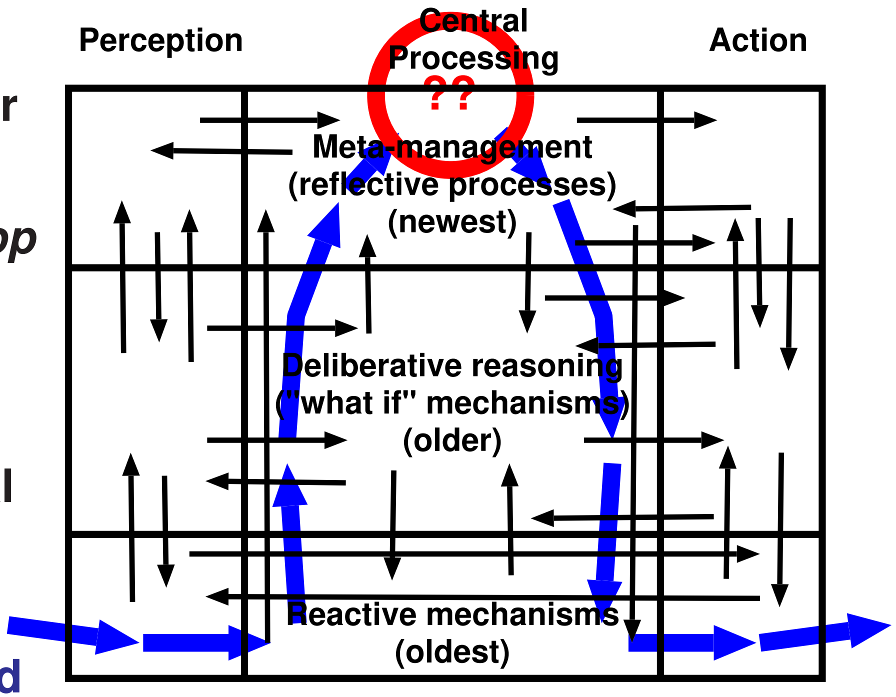
The "Omega" model of information flow

CogAff allows many variants, e.g. the "contention scheduling" model (Cooper and Shallice 2000).

Some authors propose a "will" at the top of the omega
(E.g. Albus 1981)

Rejects layered concurrent perceptual and action towers separate from central tower.

What is the difference between processes in the perceptual column and processes in the central column?



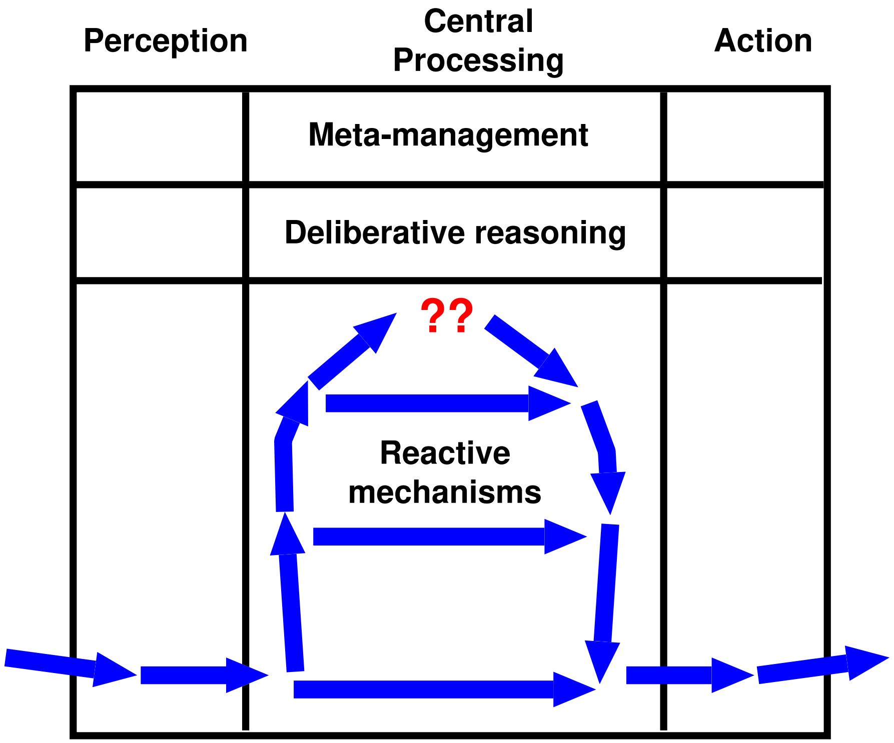
TENTATIVE ANSWER: Multi-level (multi-window) perception uses dedicated concurrent parsing and interpretation of sensory arrays, e.g. building new data-structures in registration with sensory arrays, e.g. in registration with a 2-D visual array.

Contrast "peephole" perception.

Likewise multi-window vs peephole action.

Another special case of CogAff: Subsumption architectures (Brooks)

These allow different architectural layers, but only within the reactive sub-space, where they form a sort of dominance hierarchy (unlike the layers in H-Cogaff described later.)
Brooksians deny that animals (even humans) use deliberative mechanisms.
(How do they get to overseas conferences?)



These reactive subsumption architectures are able to meet requirements for human-like capabilities ONLY IF quite unrealistic assumptions are made about evolutionary developments, storage capabilities, etc.

Subsumption and CogAff

Subsumption, like the Omega architecture and many other architectures, uses only a subset of the mechanisms allowed in the CogAff schema.

We should avoid all dogmatism and ideology, and investigate which subsets are useful for which organisms or machines, and how they might have evolved.

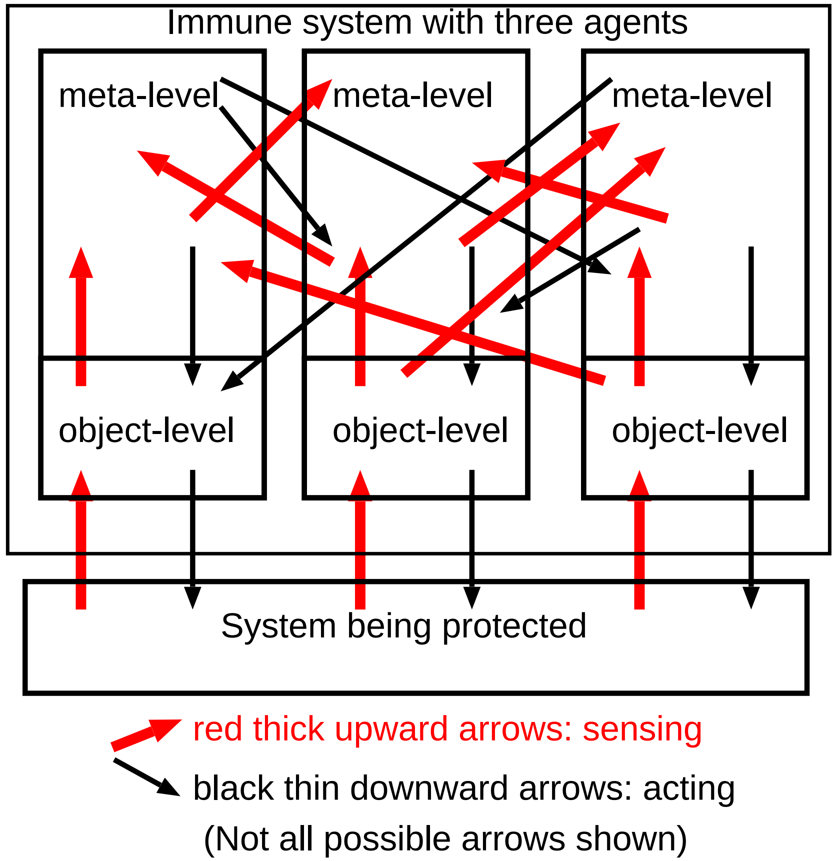
That way we'll learn instead of fighting.

A mutual meta-management system

Catriona Kennedy has been working on extending these ideas in the design of a robust system for detecting and repairing code damaged by hostile intruders.

To avoid the fragility of having only one monitor, Kennedy proposes a collection of them each observing not only the system being protected but also one another's observations, and, if appropriate, taking "corrective" action, e.g. repairing damaged code.

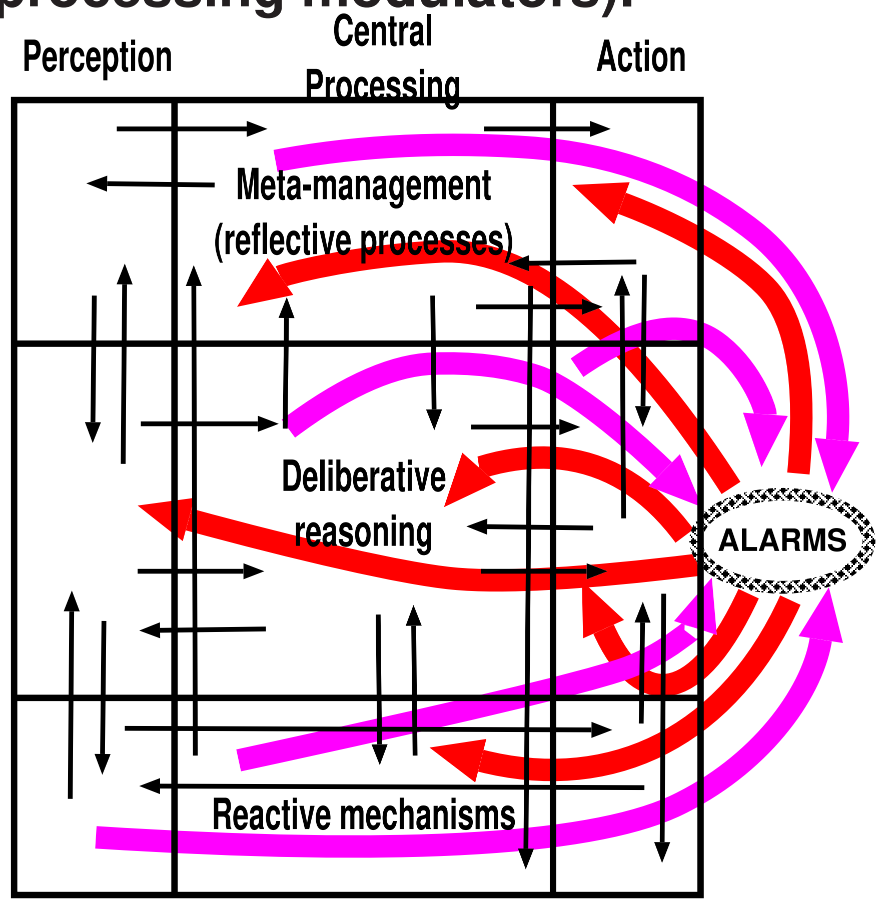
The "object level" components monitor and act on the system being protected. The meta-level components monitor and act on the object- and meta-level components (which may be reactive, deliberative or a mixture).



Some of Kennedy's papers outlining the theoretical ideas and describing a prototype implementation can be found here: http://www.cs.bham.ac.uk/research/cogaff/0-INDEX00-05.html

The need for “alarm” mechanisms

As processing grows more sophisticated, so it can be come slower, to the point of danger. A possible remedy is to use one or more fast, powerful, “global alarm systems” (processing modulators).



ALARM MECHANISMS MUST USE FAST PATTERN-RECOGNITION AND WILL THEREFORE INEVITABLY BE STUPID, AND CAPABLE OF ERROR!

Note: An alarm mechanism is just part of the reactive sub-system. Drawing it separately merely serves the pedagogic function of indicating the role.

Many variants possible. E.g. purely innate, or trainable.

E.g. one alarm system or several? (Brain stem, limbic system, ...???)

Various kinds of more or less global, more or less rapid, re-direction or re-organisation of processing.

The five Fs: Feeding, fighting, fleeing, freezing, and reproduction

(Usually only four are specified!)

Many sorts of alarms

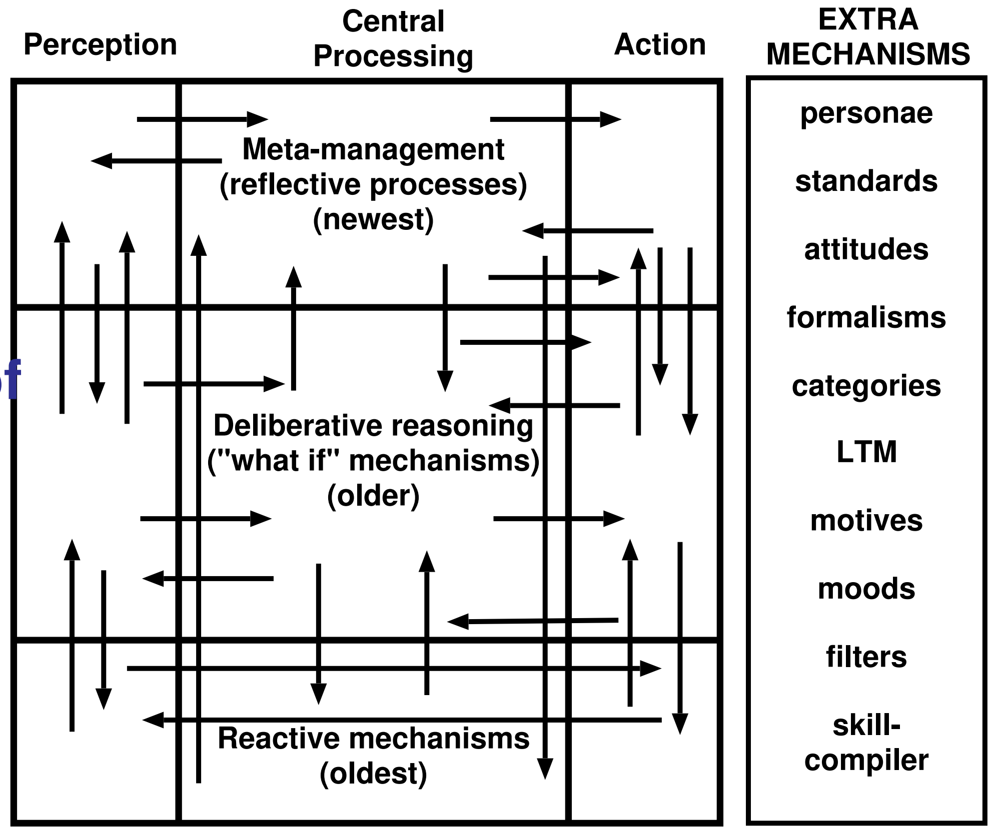
- Alarms allow rapid redirection of the whole system or specific parts of the system required for a particular task (e.g. blinking to protect eyes.)
- The alarms can include specialised learnt responses: switching modes of thinking after noticing a potential problem.
- E.g. doing mathematics, you suddenly notice a new opportunity and switch direction. Maybe this uses an evolved version of a very old alarm mechanism.
- The need for (POSSIBLY RAPID) pattern-directed re-direction by meta-management is often confused with the need for emotions e.g. by Damasio, et. al.
- Towards a science of affect:

  Not just alarms – many sorts of control mechanisms, evaluators, modulators, mood controllers, personality selectors, etc.

Additional components are needed

A partial list is on the right:

Many profound implications regarding varieties of possible architectures, possible types of learning and development, possible effects of brain damage, varieties of affective control states.



Example:

Different sorts of learning can occur within individual sub-systems and also different sorts of links between sub-systems can be learnt.

(Not only links shown so far. E.g. learning athletic skills.)

Some forms of development may ‘grow’ new subsystems.
E.g. learning to talk? Learning mathematics? Learning to paint, or play a violin, or compose music? New forms of self-control?

Varieties of motivational sub-mechanisms

What is motivation?

- A type of affective control state or process, with many sub-types.
- Different types of contents, including bringing about, preserving, increasing, reducing, preventing, removing... some state of affairs.
- Motives or goals can be short term, long term, permanent.
- They can be triggered by physiology, by percepts, by deliberative processes, by meta-management.
- They can be implicit in the operation of active mechanisms, or explicit.
- They can operate in a totally innate (genetically determined fashion) or be learnt, or influenced by a culture (e.g. whether you enjoy eating grubs).
- They can be part of the reactive system, part of the deliberative system, part of meta-management.
- They can be implicit or explicit.
- They can use a wide range of representational formalisms (e.g. with or without compositional semantics).

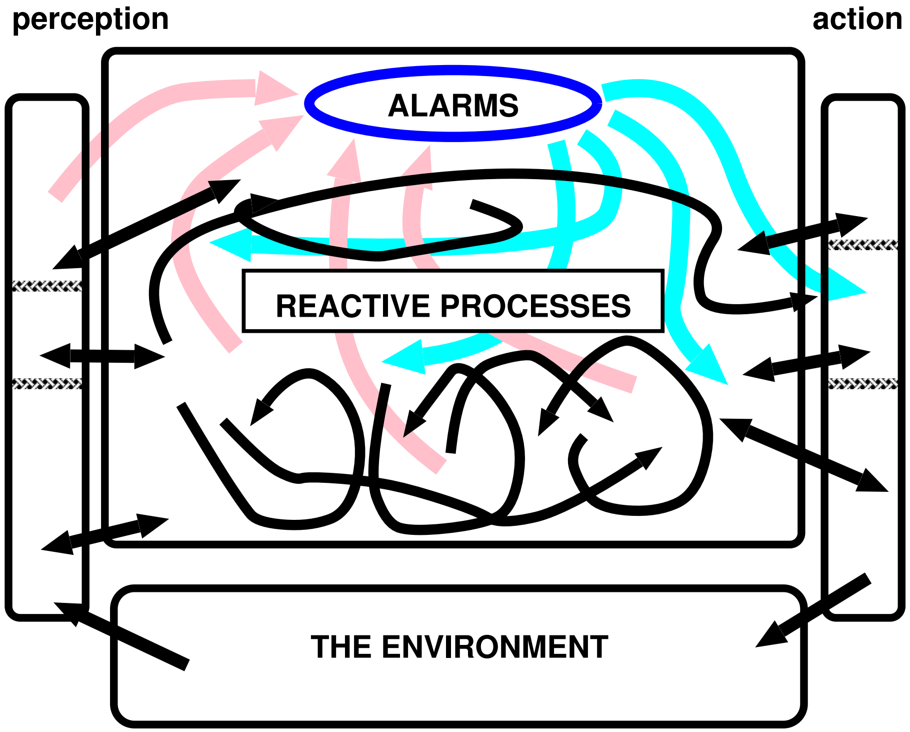
We need a better overview of the requirements for different sorts of motivational mechanisms and the requirements generated by that variety, e.g. for mechanisms that detect and resolve conflicts.

Not all parts of the grid are present in all animals: e.g. insects?

Not all organisms, and certainly not all useful robots will have all the components allowed by the CogAff schema. Consider how to design an insect including an alarm mechanism?

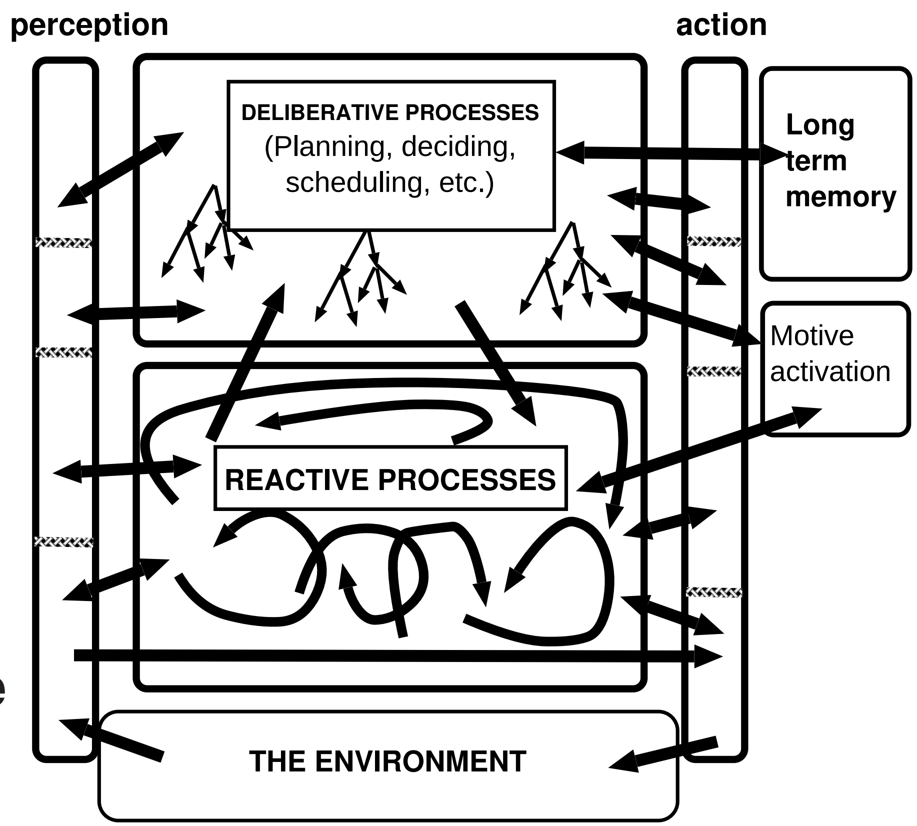
Even reactive systems may require perceptual mechanisms to operate at different levels of abstraction, e.g. recognising food, mates, danger.

There may also be hierarchical action subsystems.



Towards deliberative systems

Add a deliberative layer, e.g. for a monkey, or chimp?



The requirements of a deliberative later could form a niche applying pressure for evolution of more abstract levels of perceptual processing, e.g. chunking perceptual inputs into forms useful for learning predictive associations, or for learning which actions do what.

I.e. perception evolves to support the needs of 'what if' reasoning mechanisms.

We need to understand the varieties of deliberative mechanisms, and the forms of representation they require, from the simplest that do not need compositional semantics, to the sophisticated ones that do.

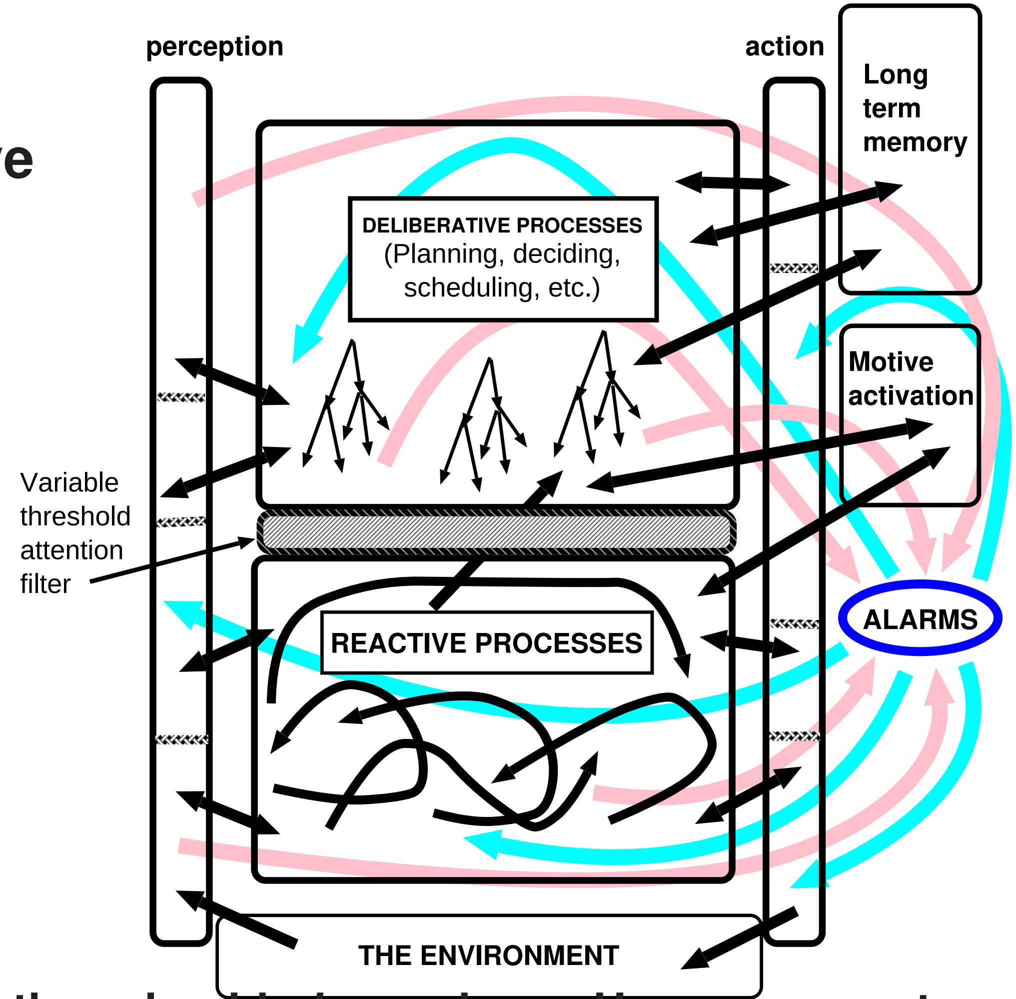
Alarm mechanism (global interrupt/override):

- Allows rapid redirection of the whole system
- sudden dangers
- sudden opportunities
- Freezing
- Fighting, attacking
- Feeding (pouncing)
- General arousal and alertness (attending, vigilance)
- Fleeing
- Mating
- More specific trained and innate automatic responses

What Damasio and Picard call “Primary Emotions” seem to be certain states generated in reactive mechanisms via global alarm systems.

Reactive and deliberative layers with alarms

Deliberative mechanisms come in various forms. The most sophisticated ones have complex architectural requirements, indicated only sketchily above.



What Damasio and Picard call "Secondary Emotions" seem to be reactions triggered by central cognitive processes in a deliberative mechanism.

Note: Whether these involve the same physiological responses as primary emotions in humans and other animals is an empirical question.

There is no *theoretical* reason why they should *always* do so. Humans seem to vary in this respect — e.g. in how grief and joy affect them. There seems to be an attention filter with dynamically varying threshold. E.g. pain can be temporarily suppressed or ignored when there are urgent and important tasks.

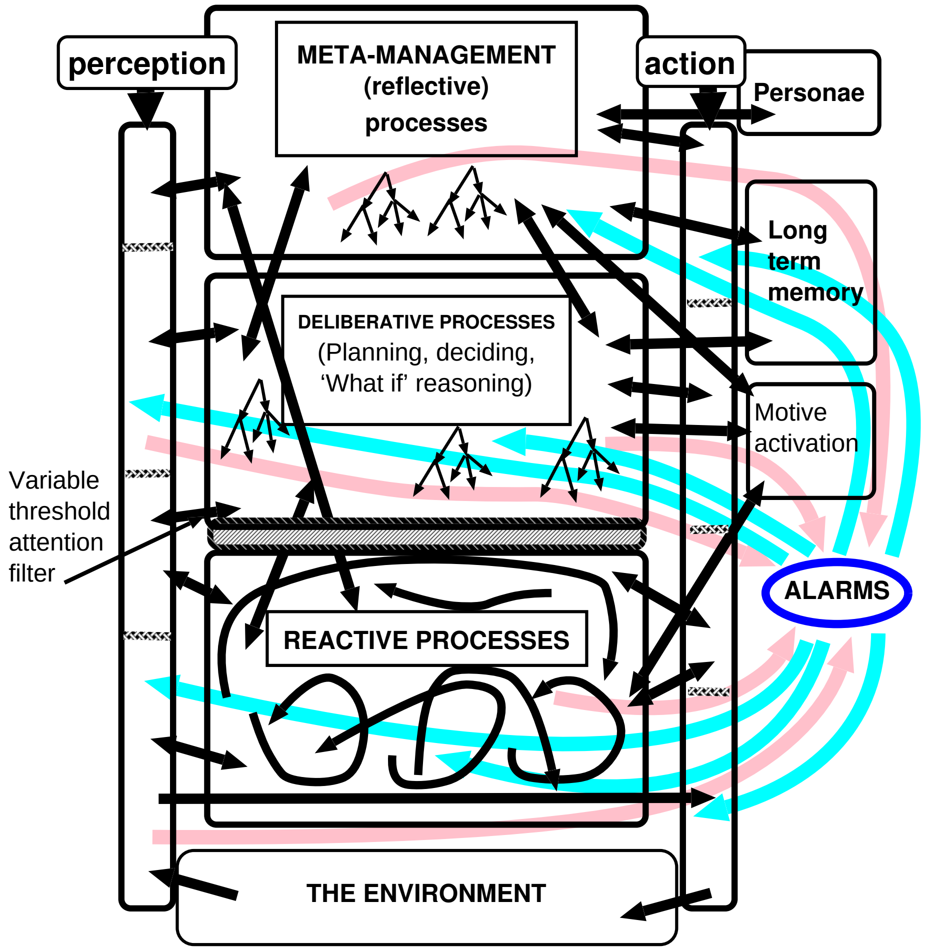
H-COGAFF: A human-like architecture.

An instance of the CogAff schema using all the components.

The diagram is very impressionistic, not a precise "blue-print".
Described in more detail in papers in the Cogaff directory:
http://www.cs.bham.ac.uk/research/cogaff/

Probably includes several alarm mechanisms. (Brain stem, limbic system, blinking reflexes, ...???)

Attention filter is needed to protect resource-limited deliberative and meta-management systems from relatively unimportant interrupts from reactive and alarm mechanisms.



But no filter is perfect, and some emotional states come from low importance interrupts given "high insistence" by stupid alarm systems: a type of "perturbance" (tertiary emotion).

Tertiary emotions

(Called “perturbances” in older Cogaff project papers.)

- Involve interruption and diversion of thought processes.

  I.e. the meta-management layer does not have complete control.
- Question: Is it essential that all sorts of emotions have physiological effects outside the brain, e.g. as suggested by William James?

  No: which do and which do not is an empirical question, and there may be considerable individual differences.
- An organism that does not have meta-management cannot control attention, etc. and therefore cannot LOSE that sort of control, and therefore cannot have tertiary emotions.
- It does NOT follow that tertiary emotions are required for intelligent control.

  (Damasio’s non-sequitur.)

Different architectural layers support different sorts of mental states and processes.

- We can use the layers to define *architecture-based ontologies for different sorts of minds*.
- *Describing different animals will require using different mental ontologies*
- *Humans at different stages of development will instantiate different mental ontologies.*

Some notes:

- Different aspects of love, hate, jealousy, pride, ambition, embarrassment, grief, infatuation can be found in all three categories of emotions.
- Remember that these are not *static* states but *developing* processes, with very varied aetiology. Different patterns of growth and decay correspond to different sorts of emotions.
- We don't necessarily already have names for all the significantly different cases
- Not all emotions are necessarily useful. Some can be seriously dysfunctional.
- Moods are global control states often confused with emotions.
- Attitudes (e.g. love of one's country) are specific cognitive states often confused with emotions.

Socially important human emotions

- These involve rich concepts and knowledge and high level control mechanisms (architectures).
- Some emotions use categories for self-description and self-evaluation that are absorbed from a culture.
- Some socially important processes involve switching between different personalities in different social contexts.

Example: longing for someone or something:

- Semantics: To long for something you need to know of its existence, its remoteness, and the possibility of being together again.
- Control: One who has deep longing for X does not merely occasionally think it would be wonderful to be with X. In deep longing thoughts are often *uncontrollably* drawn to X. Moreover, such longing may impact on various kinds of high level decision making as well as the focus of attention.

  Physiological processes (outside the brain) may or may not be involved. Their importance is over-stressed by some experimental psychologists.

Brains support consciousness? How?

What's consciousness?

People assume consciousness is one thing.

Then they ask questions like:

- **which animals have IT?**
- **how did IT evolve?**
- **what is ITS function?**
- **could machines have IT?**
- **which bits of the brain produce IT?**

If there’s no “IT” the questions make no sense.

- What we call “consciousness” is a large ill-defined COLLECTION of capabilities.
- THEY (the various capabilities) can be present or absent in different combinations, in different animals, in people at different stages of development or after brain damage.
  Also in different machines.
- No pre-ordained subset of that set of capabilities is THE subset required for consciousness.
  - Compare flea, fish and frog consciousness.
  - Compare infant and adult human consciousness.
- Not just ONE thing that is always present or absent. Neither is it a matter of degree.
- I.E. “CONSCIOUSNESS” IS A PARTLY INDETERMINATE “CLUSTER CONCEPT”.
  (Like “emotion”)
- People think they know what IT is from experience.
  Before Einstein people thought they knew what simultaneity was from experience. We can unintentionally fool ourselves.

(The notion of a “cluster concept” is explained briefly in this slide presentation: http://www.cs.bham.ac.uk/research/cogaff/ibm02/)

Varieties of consciousness

By exploring varieties of awareness of the environment and varieties of self-awareness made possible by different architectures we can distinguish different varieties of consciousness.

- **Microbe consciousness**
- **Flea consciousness**
- **Frog consciousness**
- **Eagle consciousness**
- **Chimp consciousness**
- **Infant (human) consciousness**
- **Adult consciousness**
- **Varieties of drug-modified consciousness**

See talk 9 here http://www.cs.bham.ac.uk/˜axs/misc/talks/ (on varieties of consciousness.)

Towards a conclusion ....

Understanding what human beings are, and being able to design an implementable human-like robot requires us, at least, to understand:

– the varieties of affordances in human environments, and ways in which affordances can be perceived, represented, and used in acting in the environment.
– the varieties of types of co-evolved concurrently active information-processing sub-systems that make up a human being: the CogAff schema provides only a first draft, coarse-grained, taxonomy.
– how all these subsystems develop, and how they interact with one another, and what sorts of states they can generate (e.g. varieties of emotions, moods, pleasures, pains, and other affective states).
– the varieties of forms of representation deployed within the different subsystems, and how they are used in learning various ontologies that develop during a person's life.
– the varieties of ways in which the system can go wrong, producing both genetic malfunctions and manifestations of various kinds of brain damage and disease. (E.g. R.Barkley on ADHD)
– the capabilities that appear to exist only in humans, and why they do not occur in other animals. This includes understanding how much of a typical human mind is genetically programmed, how much a social product, etc.

Conclusion:
We need more focus on requirements

In short: we need to clarify the requirements to be satisfied by a design for a human-like mind before we can hope to produce such a design. Testing partial designs can, as AI has shown, feed into refining the requirements for extending those designs.

There's obviously a lot more work to be done (300 years? 3000?)

THE PROJECT NEEDS A LOT MORE RESEARCHERS, FROM MANY DIFFERENT DISCIPLINES.

PLEASE JOIN IN.

More Acknowledgements

There is considerable overlap with ideas about architectures in the work of Marvin Minsky, e.g. in *The Society of Mind* and in his draft book *The Emotion Machine* available on his web site:
http://web.media.mit.edu/˜minsky/

There is also overlap with John McCarthy's papers
http://www-formal.stanford.edu/jmc/

Compare Dennett's book *Kinds of minds*

NO DOUBT THERE ARE OTHER RELEVANT PUBLICATIONS.

Some related web sites

The Birmingham Cognition and Affect Project

PAPERS (mostly postscript and PDF):
http://www.cs.bham.ac.uk/research/cogaff/
(References to other work can be found in papers in this directory)

TOOLS:
http://www.cs.bham.ac.uk/research/poplog/freepoplog.html
(Including the SIM_AGENT toolkit)

SLIDES FOR TALKS (Including IJCAI01 philosophy of AI tutorial with Matthias Scheutz):
http://www.cs.bham.ac.uk/~axs/misc/talks/

Free online book: The Computer Revolution in Philosophy (1978)
http://www.cs.bham.ac.uk/research/cogaff/crp/
(With some recently added notes and comments.)

SEI at CMU

Related information and discussion, and a list of definitions of “software architecture” can be found here: http://www.sei.cmu.edu/ata/ata_init.html