THE UNIVERSITY
OF BIRMINGHAM

A SYSTEMS APPROACH TO CONSCIOUSNESS

Aaron Sloman

School of Computer Science
Cognitive Science Research Centre
Email: A.Sloman@cs.bham.ac.uk
WWW: http://www.cs.bham.ac.uk/˜axs

WE NEED TO DISTINGUISH:

- **Empirical questions**
- **Design questions**
- **Conceptual questions**

THIS TALK IS MAINLY ABOUT CONCEPTUAL QUESTIONS AND DESIGN QUESTIONS.

We need to sort out conceptual questions in order to know:

- **What we are trying to explain or build**
- **Whether we are making progress**
- **When we are arguing at cross purposes**

CONFUSIONS ABOUT CONSCIOUSNESS

We use words like ‘conscious’, ‘aware’, ‘experience’:

- **as if they had clear fixed meanings,**
- **as if there were a binary division between things they apply to and things they do not apply to.**

This generates the illusion that *consciousness* **is something that is either present or absent in an object.**

THIS TEMPTS US TO ASK PSEUDO-QUESTIONS:

- **Which animals have consciousness?**
- **How did it evolve?**
- **Does it have a biological function?**
- **Is it reducible to physics?**
- **Could a robot have it?**
- **Could there be a machine (a “zombie”) with all the external appearance of having consciousness, but without having it?**

There are also lots of muddles about “qualia”, which I have no time to discuss.

EVIDENCE FOR CONCEPTUAL CONFUSION

CONTRADICTIONS IN OUR ORDINARY WAYS OF THINKING ABOUT CONSCIOUSNESS:

- **Are you conscious when terrified in a dream?**
- **If you first become aware of a noise when it stops, were you conscious of it before it stopped?**
- **Is a sleep-walker who dresses himself and walks down a staircase conscious?**

THERE ARE ALSO VERY UNCLEAR BOUNDARIES:

- **When a foetus develops, when does it become conscious?**
- **Where can we draw the line between animals with and animals without consciousness?**
- **When a degenerative brain disease gradually reduces a normal person to an apparent vegetable, at what point does consciousness disappear?**
- **Some forms of brain damage produce “blindsight” – people claim not to be able to see, and yet they can answer questions about where a light is.**

SOME FAMILIAR DEMONSTRATIONS

There are examples where it is not clear whether one is or is not conscious of something.

```
          / \
         / A \
        /     \
       / BIRD  \
      /         \
     /  IN  THE  \
    /             \
   / THE    HAND   \
  /_________________\
```

Some people see only a familiar phrase? Do you?

(If you see something wrong with the text, please don’t inform your neighbour.)

If the audience includes appropriate “subjects” I’ll perform an experiment.

WHAT'S IN THE BLIND SPOT?

Another case that causes confusion is how to describe what is happening at the 'blind spot'.

Do we, or don't we, experience something there?

X **O**

Look at the "X" with the left eye closed, and move back and forth.

At a certain distance the "O" disappears.

So, when we shut one eye, why don't we see a gap where the blind spot is?

Compare looking at the left hand edge of a page of text, while the left eye is shut. Move the page back and forth. Which words on the right disappear?

What do we see?

Shut one eye and look around you: where's the gap in what you see?

IS THERE A GAP OR ISN'T THERE?

DOES YOUR VISUAL FIELD HAVE BOUNDARIES?

CONTENTS OF PERIPHERAL VISION ARE EXTREMELY UNCLEAR TO US

NO DICHOTOMY and NO CONTINUUM

A TEMPTING MISTAKE

It is often suggested (e.g. by Susan Greenfield in last week's lecture) that we can avoid the paradoxes by thinking of consciousness as a matter of *DEGREE*. **So:**

- **Differences between animals are differences of degree.**
- **Differences between states of consciousness are matters of degree.**
- **Differences in brain mechanisms and brain states are matters of degree.**

Though there's some truth in this, I believe that it is a serious over-simplification.

WHAT ALTERNATIVES ARE THERE TO THESE TWO INADEQUATE VIEWS:

- **Consciousness is a clearly defined state that is always either present or absent.**
- **Consciousness is a matter of degree, and subject to quantitative variation.**

ANSWER:

WE ARE TALKING ABOUT CLUSTER CONCEPTS

CONSCIOUSNESS A “CLUSTER” OF RECOMBINANT CAPABILITIES

While apparently talking about *ONE* thing we may be talking about a very complex *CLUSTER* of different things.

DIFFERENT SUBSETS OF THE CLUSTER OCCUR:

- **in different organisms,**
- **in different machines,**
- **in different people,**
- **even in the same person at different times:**

infancy,
childhood,
adulthood,
during senile dementia,
after brain injury,
and so on.

NB:

Lots of discontinuous changes may come close to a smooth continuum, but we need to understand the discontinuities.

THERE IS NO “UNIQUE” SUBSET

OF CAPABILITIES THAT DEFINES “CONSCIOUSNESS”

It’s not a disjunction either.

WHAT CAPABILITIES?

There is a vast collection.

- **Many different kinds of perceptual capabilities. (Contrast recognition, interpretation, grasping structure, seeing possibilities, controlling posture or motion.)**
- **Different kinds of memories: Short term buffers, of varying length. Long term associative memory. Longer term storage supports both cognitive maps and plan creation.**
- **Many different kinds of learning (including new concepts, new languages, rules, new motor skills, learning to recite poems, learning a complex dance, learning to perform a Beethoven piano sonata).**
- **Many different kinds of motivational processes: (Common biological drives, plus curiosity, aesthetic desires, long term goals, suspended plans, “anytime” planning, moral feelings, ideals, socially acquired tastes, etc.)**
- **Different kinds of self-monitoring, self-evaluation, self-control, attending to current internal states.**
- **The ability to control thought processes, or to lose control (e.g. in emotional states like humiliation, guilt, infatuation, obsession.)**
- **The ability to think about mental processes in others.**

DESIGNS AND NICHES

DIFFERENT COMBINATIONS OF CAPABILITIES CORRESPOND TO DIFFERENT DESIGNS

- **The task of an engineer is typically to create a design that will satisfy (or come close to satisfying) a combination of requirements and constraints.**
- **A design is an integrated collection of capabilities linked together in an *IMPLEMENTATION*.**
- **Evolution can be seen as producing designs: though there is no designer or engineer, only natural selection.**
- **Biologists use the notion of a "niche" to talk about the set of requirements and constraints: i.e. what a design satisfies, more or less well. (A niche is an abstraction, not a geographical region.)**

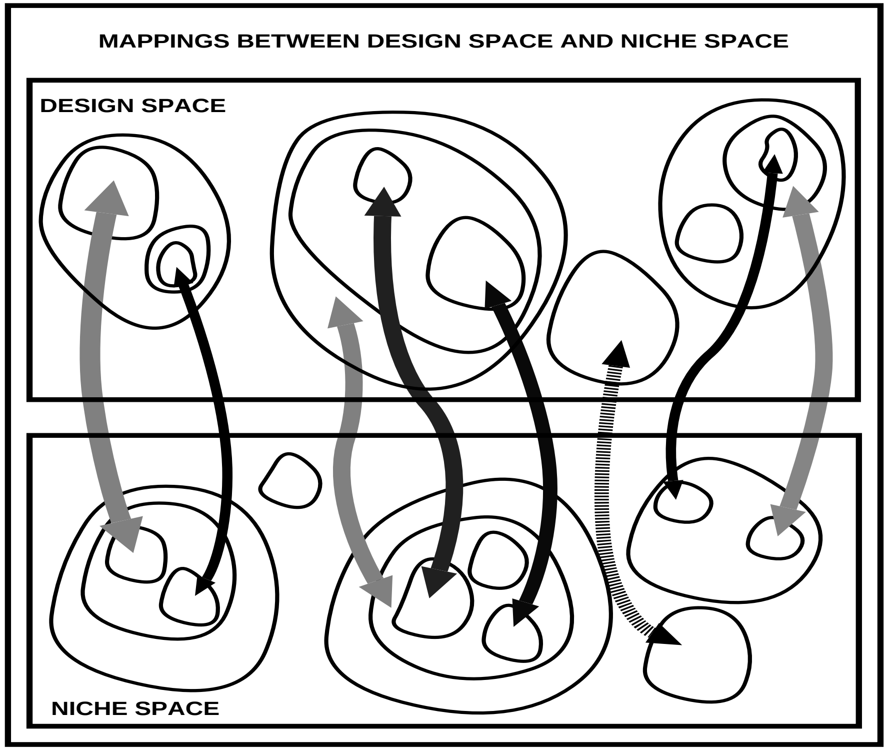
DIFFERENT SORTS OF REQUIREMENTS CORRESPOND TO DIFFERENT NICHES

AI IS THE GENERAL STUDY OF DESIGN SPACE, NICHE SPACE AND THEIR INTERRELATIONS.

AI USES COMPUTERS, BUT COULD, IN PRINCIPLE, USE OTHER MECHANISMS: HYBRID DESIGNS ARE IMPORTANT.

CONJECTURE:
Architecture is more important than mechanism.

DESIGN SPACE and NICHE SPACE



NOTES

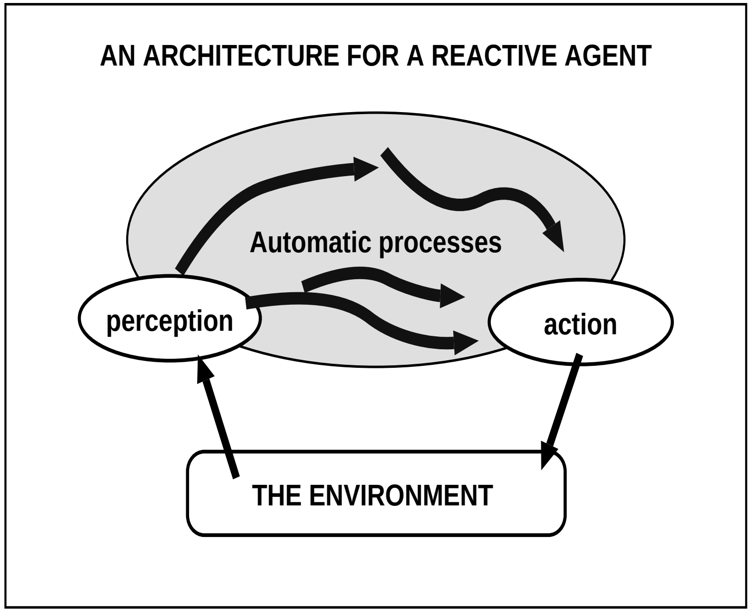
- **A niche is a set of requirements**
- **A design is a set of specifications**
- **Mappings are not unique: there are always trade-offs**

DYNAMICS:

WE NEED TO UNDERSTAND TRAJECTORIES

- **Possible within an agent (development, learning)**
- **Possible only across generations (evolution, ALIFE)**

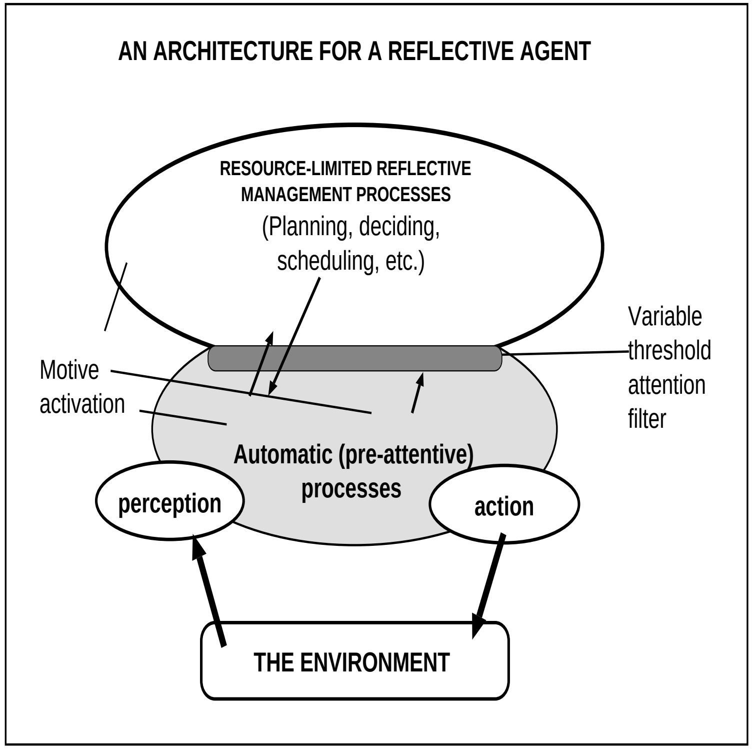
REACTIVE AGENTS



IN A REACTIVE AGENT:

- **Mechanisms and space are pre-allocated to specific tasks**
- **There is no construction of new plans**
- **There is no explicit evaluation of alternative plans**
- **Parallelism gives speed**
- **There may be tunable control loops**
- **The agent can survive even if it has only genetically determined behaviours**
- **Difficulties arise if the environment requires new plan structures.**

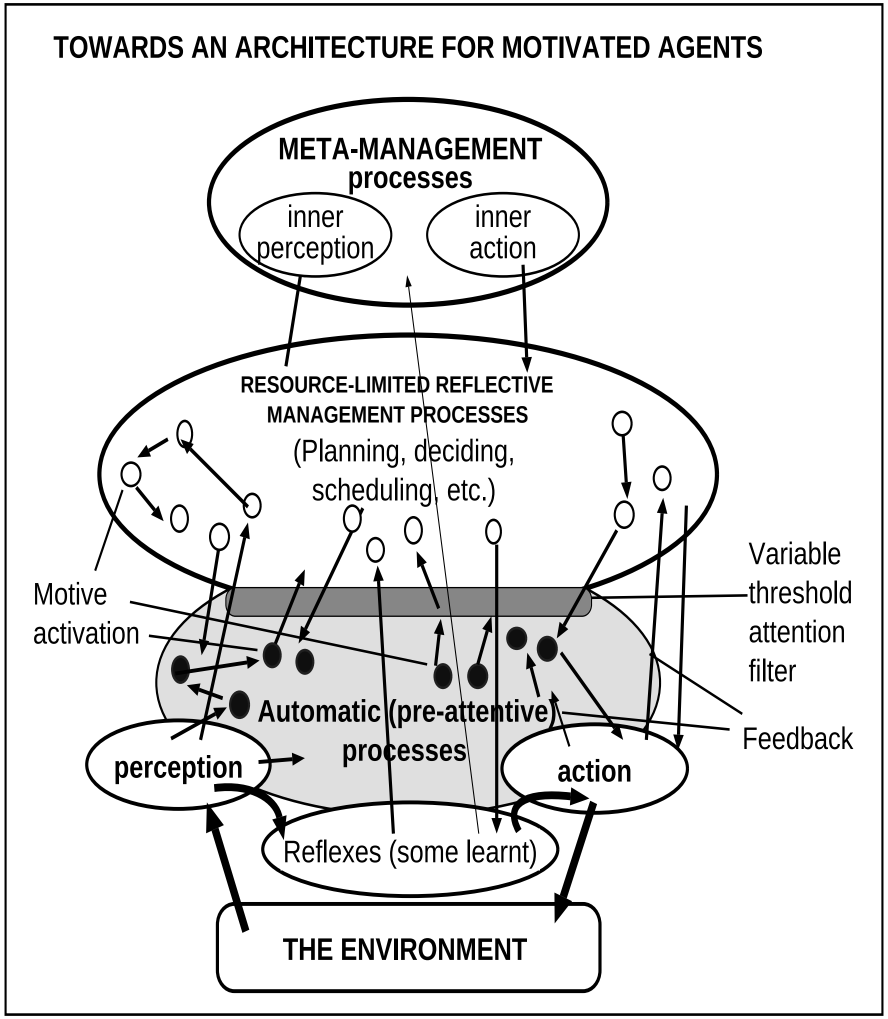
TOWARDS REFLECTIVE AGENTS


AN ARCHITECTURE FOR A REFLECTIVE AGENT
RESOURCE-LIMITED REFLECTIVE MANAGEMENT PROCESSES
(Planning, deciding, scheduling, etc.)
Variable threshold attention filter
Motive activation
Automatic (pre-attentive) processes
perception
action
THE ENVIRONMENT

A REFLECTIVE AGENT

- **Mechanisms and space are dynamically allocated**
- **New plans may be constructed**
- **Options are explicitly evaluated before selection**
- **Parallelism is much reduced (for various reasons):**
  - **Learning**
  - **Access to associative memory**
  - **Integrated control**
- **A fast changing environment can cause too many interrupts, frequent re-directions.**

TOWARDS AUTONOMOUS AGENTS



TOWARDS AN ARCHITECTURE FOR MOTIVATED AGENTS

META-MANAGEMENT processes
inner perception
inner action
RESOURCE-LIMITED REFLECTIVE MANAGEMENT PROCESSES
(Planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
Automatic (pre-attentive) processes
perception
action
Feedback
Reflexes (some learnt)
THE ENVIRONMENT

Towards an architecture for an autonomous agent

- **Meta-management controls contention in management processes**
- **Global monitoring can support ‘self evaluation’**

PERCEPTION CAN USE AN INTRICATE ARCHITECTURE

Perception is not just a matter of registering or recognising.

It also involves:

- **Classification at different levels of abstraction: a square, a rectangle, a quadrilateral, a polygon, a figure.**
- **Interpretation: mapping from one domain to another. E.g. the 2-D optic array is interpreted in terms of a 3-D environment. Acoustic patterns are interpreted as meaningful speech.**
- **Grasping structure: seeing not only eyes, nose, mouth, arms, legs, hands, feet, but how they are related together. The hands are on the ends of the arms, but a finger may be touching the nose.**
- **Grasping patterns of change and motion: the wasp is flying towards the window, the car is moving forwards while its wheels are turning, the scissors are opening and shutting.**

Perception Continued

- **Grasping possibilities and constraints inherent in structure (what J J Gibson called “affordances”: a chair can support you, a table can obstruct motion, a door allows transfer to another room a window catch allows the window to be held open, a handle allows an object to be grasped.**

Thus a human-like (or ape-like?) perceptual system needs to be able to create and manipulate

- **a number of different sorts of rapidly changing representations**
  - **of different sorts of information,**
  - **using:**
    - **incoming data,**
    - **prior knowledge,**
    - **current motivation**

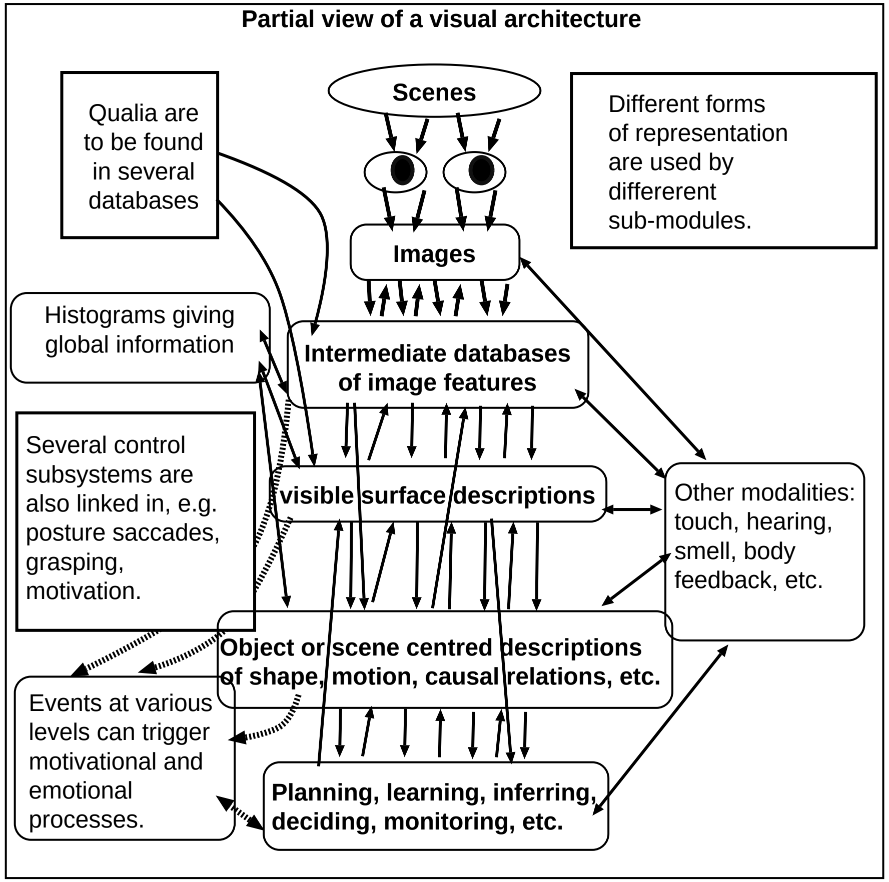
Human perceptual architectures also allow the agent to attend to some aspects of these *INTERNAL* information stores.

E.g. learning to draw, sighting a gun.

This is one of the sources of concerns about “qualia”.

BUT OUR ACCESS IS BOTH INCOMPLETE AND UNRELIABLE!

DESIGNING A VISUAL SYSTEM



Partial view of a visual architecture

Qualia are to be found in several databases

Scenes

Different forms of representation are used by differerent sub-modules.

Images

Histograms giving global information

Intermediate databases of image features

Several control subsystems are also linked in, e.g. posture saccades, grasping, motivation.

visible surface descriptions

Other modalities: touch, hearing, smell, body feedback, etc.

Object or scene centred descriptions of shape, motion, causal relations, etc.

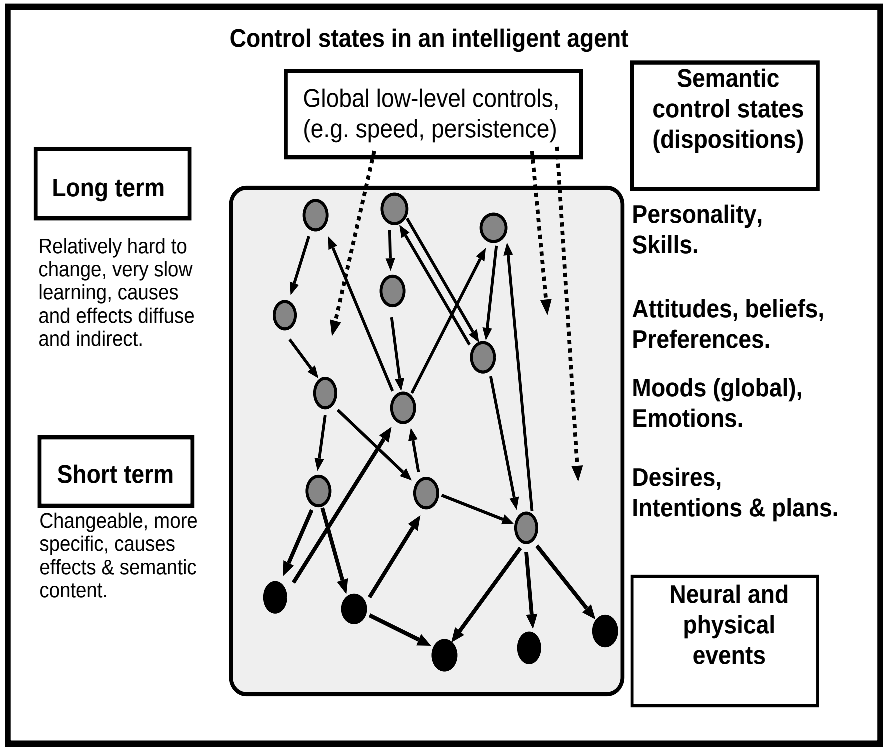
Events at various levels can trigger motivational and emotional processes.

Planning, learning, inferring, deciding, monitoring, etc.

Towards an architecture for a visual system

- **There are many intermediate information structures.**
- **Higher level processes may be able to access them.**
- **Reflecting on them gives rise to questions about experiences, qualia, etc.**
- **This could happen in robots.**

TYPES OF CONTROL STATES



Control states in an intelligent agent

Global low-level controls, (e.g. speed, persistence)

Semantic control states (dispositions)

Long term

Relatively hard to change, very slow learning, causes and effects diffuse and indirect.

Short term

Changeable, more specific, causes effects & semantic content.

Personality, Skills.

Attitudes, beliefs, Preferences.

Moods (global), Emotions.

Desires, Intentions & plans.

Neural and physical events

Control states of varying scope and duration

The "higher" states are:

● Harder to change

● More long lasting

● Subject to more influences

● More general in their effects

● More indirect in their effects

● More likely to be genetically determined(??)

ARCHITECTURALLY GROUNDED CONCEPTS

We can replace endless debates at cross-purposes with research that makes real progress, in philosophy and in science.

- **A design specifies an architecture.**
- **The architecture supports a variety of states and processes.**
- **Analysis of possible states and processes generates families of theory-based concepts.**
- **These new concepts can elaborate and extend common sense concepts, as happened when physics gave us a new architecture for matter.**
- **The new concepts enable us to ask new questions: not**
  - **Which animals are conscious? but**
  - **Which kinds of consciousness do different animals have?**
- **It's not enough just to understand one architecture, or to build one type of robot. Deep understanding requires us to explore regions and trajectories in design space and niche space.**

WARNING:

THE PROBLEMS ARE VERY HARD
AND PROGRESS WILL BE SLOW

IMPLICATIONS

- **Such an architecture can provide a basis for a deeper understanding of how the human mind normally works, and how it might go wrong, helping therapy, counselling and education.**
- **A design-based theory can generate a host of new empirical questions to be settled by neurophysiological, psychological and biological research.**
- **Many designs involve creation of "virtual" machines, e.g. word-processors, compilers, operating systems. These are information processing machines that operate on abstract entities. But they can have real causal powers, and form part of a control system, e.g. for a factory or aeroplane.**
- **Understanding virtual machines and how they relate to the mechansims in which they are** *IMPLEMENTED* **is an important task for philosophy.**
- **We may find that certain high level aspects of a human-like architecture can be implemented on quite different sorts of low level mechanisms (e.g. computer-based mechanisms).**

CONJECTURE

Architecture dominates mechanism

Acknowledgements

This work was supported by the UK Joint Council Initiative, the Renaissance Trust, and a grant from the University of Birmingham.

Some of the work is now funded by DRA Malvern

Thanks to various members of the Cognition and Affect project at the University of Birmingham:

Luc Beaudoin, Ian Wright, Tim Read, Christian Paterson, Chris Complin, Riccardo Poli, Darryl Davis, Brian Logan,

Papers (mainly compressed postscript) by the group can be found at the ftp site:

ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/

See also

WWW: http://www.cs.bham.ac.uk/˜axs/

and the misc/ sub-directory.