WHAT SORT OF ARCHITECTURE CAN SUPPORT EMOTIONALITY?

(MIT Media Lab 1996)

Aaron Sloman
School of Computer Science
The University of Birmingham
Email: A.Sloman@cs.bham.ac.uk
http://www.cs.bham.ac.uk/~axs

 objetoives:

• A mythical history of the evolution of human-like minds
• Different architectures support different sorts of emotions: found in different animals
• Some speculations about architectures for human-like machines.
• Various sub-architectures in human brains support very different types of emotions (e.g. based on “reactive” “deliberative” and “meta-management” layers).
• Hence the ENORMOUS confusion in psychological, philosophical and neuroscience literature. (People argue at cross-purposes)
• Comments on characteristically human emotional states involving partial loss of control of thought processes.
• A brief characterisation of Artificial Intelligence as exploration of “design space” and “niche space” and trajectories therein.

(See also

http://www.cs.bham.ac.uk/~axs/
and the misc/ sub-directory.)
KINDS OF QUESTIONS

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

Design questions are concerned with:

- What sorts of designs are possible?
- What they are good for (what niche)?
- How different designs differ?
- How they can be implemented?
- Which concepts of types of mental states they support. (That’s partly conceptual)

CONFUSIONS TO AVOID: ASKING THE WRONG QUESTIONS

Minsky: “dumbbell” questions, presupposing dichotomies, are often misguided.

WHICH THINGS ARE CONSCIOUS AND WHICH ARE NOT?

It is not a dichotomy:

\[
\text{THINGS WITH MINDS} \quad \text{THINGS WITHOUT MINDS}
\]

It is not a continuum:

virus … amoeba …… flea …… mouse …… monkey …… bonobo …… human … ??

THERE ARE MANY DISCONTINUITIES
WE NEED TO STUDY THEM!

DON’T ASSUME IT’S A CONTINUUM

Lots of discontinuous changes may come close to a smooth continuum, but we need to understand the DIScontinuities in “design space”.
MENTALITY: A “CLUSTER” OF RECOMBINABLE CAPABILITIES

What is it to have a mind? To have emotions?
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....

NO UNIQUE SUBSET OF CAPABILITIES DEFINES “CONSCIOUSNESS” “INTELLIGENCE” “EMOTION”

DISCONTINUITIES OCCUR BETWEEN DIFFERENT SUBSETS.
Example:
- A rat can be afraid.
- Can it be humiliated? Ambitious? Guilt-ridden?
- WHY NOT?

LAYERS OF ARCHITECTURE IN A DESIGN

CONJECTURE:
Humans have at least three architectural layers:
- A very old reactive layer shared with many kinds of animals (insects?)
- A deliberative layer shared with some other animals
- A meta-management layer shared with very few animals.

- THE REACTIVE LAYER PROVIDES SPEED
  At the cost of rigidity.

- DELIBERATIVE LAYER PROVIDES FLEXIBILITY AND CREATIVITY
  At the cost of seriality and slowness.

- META-MANAGEMENT (SELF MONITORING) PROVIDES RESOURCE MANAGEMENT
  Using self monitoring and self control.
  But control and monitoring are never total.
  Monitoring may be partly inaccurate.

WARNING: IT’S NOT REALLY THAT SIMPLE!
IN A REACTIVE AGENT:
- Mechanisms and space are dedicated to specific tasks
- There is no construction of new plans
- There is no explicit evaluation of alternative plans
- Parallelism gives speed
- Some learning: e.g. tunable control loops
- The agent can survive even if it has only genetically determined behaviours
- Difficulties arise if the environment requires new plan structures.
- May not matter if individuals are cheap and expendable (insects?).

A DELIBERATIVE AGENT
- Mechanisms and space are dynamically allocated
- New plans may be constructed
- Options are explicitly evaluated before selection
- Learnt skills can be transferred to the reactive layer
- 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.
- Filtering via dynamically varying thresholds helps but does not solve all problems.
THE NEED FOR SELF-MONITORING (META-MANAGEMENT)

**Deliberative Mechanisms can be too rigid.**

**Internal Monitoring Mechanisms:**
- Record events, problems, decisions taken by the deliberative mechanism,
- Allow diagnosis of injuries and illness.
- Evaluate, relative to high level long term generic objectives, or standards.

**Generic Objectives could include:**
- Reducing frequency of failure in tasks
- Not allowing one goal to interfere with other goals
- Not wasting time on problems that turn out not to be solvable
- Not using a slow and resource-consuming strategy if a faster or more elegant method is available
- Detecting possibilities for structure sharing among actions.

Different kinds of meta-management are likely to be found in different animals. Probably some have none.

*Meta-meta-management may not be needed if meta-management is recursive?*

Towards an architecture for an autonomous agent
- Meta-management controls management processes
- Global monitoring can support ‘self evaluation’
Different architectural layers support different sorts of emotions:

The reactive layer supports:
- being startled
- being disgusted by horrible sights and smells
- being terrified by large fast-approaching objects?
- sexual arousal? Aesthetic arousal?

The deliberative layer supports:
- being frustrated by failure
- being relieved at avoiding danger
- being anxious about things going wrong
- being pleasantly surprised by success

The self monitoring meta-management layer, supports:
- having and losing control of thoughts and attention:
  - Feeling ashamed of oneself
  - Feeling humiliated
  - Aspects of grief, anger, excited anticipation, pride, and many more HUMAN emotions.

NOT EVERYTHING SUPPORTED BY A MECHANISM IS PART OF ITS FUNCTION: MULTI-PROCESSING OPERATING SYSTEMS SUPPORT THRASHING!

SOME FUNCTIONAL MECHANISMS HAVE DYSFUNCTIONAL CONSEQUENCES.

We can replace endless debates at cross-purposes with research that makes real progress:

- Each design specifies a class of architectures.
- We can try to find out which mechanisms can support those architectures.
- An architecture supports a variety of states and processes.
- Analysis of possible states and processes generates families of theory-based concepts.
  Compare: the periodic table of the elements.
- 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: E.g.
  - NOT which animals/machines are conscious/have emotions?
  - BUT which kinds of consciousness/emotions do different animals/machines have?
- It’s not enough just to understand ONE architecture. Deep understanding requires us to explore REGIONS and TRAJECTORIES in design space and niche space.
TYPES OF CONTROL STATES

Besides emotions there are personality, attitudes, moods, desires, wishes, intentions, etc.

Control states in an intelligent agent

- Global low-level controls, (e.g. speed, persistence)
- Semantic control states (dispositions)
- Personality, Skills.
- Attitudes, beliefs, Preferences.
- Moods (global), Emotions.
- Desires, Intentions & plans.
- Neural and physical events

Short term
Changeable, more specific, causes effects & semantic content.

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

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

DESIGNS AND NICHES

DIFFERENT COMBINATIONS OF CAPABILITIES CONSTITUTE DIFFERENT DESIGNS

DIFFERENT SORTS OF REQUIREMENTS CONSTITUTE DIFFERENT NICHES

- A design is an integrated collection of capabilities that can be linked together in an IMPLEMENTATION.
- A design is an ABSTRACTION which can have INSTANCES.
- 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.)
- Engineers assess designs in terms of satisfying (or coming close to satisfying) a combination of requirements and constraints.

Note: Artificial Intelligence 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.
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
- Designs need no designer, requirements no requirer.

DYNAMICS:
Which trajectories are possible:
- Within an agent (development, learning)?
- Across generations (evolution, ALIFE)?

The “Turing test” defines a tiny niche region .... of relatively little interest, except as a technical challenge.

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 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!

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 also happen in robots.

That’s one source of concepts about consciousness!
CONCLUSION

- Design-based studies not only help us understand normal functioning but also show how things might go wrong, helping therapy, counselling and education.

- Design-based theories can generate 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.

- That could be the source of many puzzles about minds.

- Progress may require new kinds of mathematics, for very complex systems whose dynamics is separate from physics.

- We may find that high level aspects of a human-like architecture can be implemented on varying sorts of low level mechanisms (e.g. computer-based mechanisms and others).

- We MAY need entirely new types of mechanisms to explain details, but assuming that now is premature.

- Beware of rejecting theories because you don’t like the idea of their being true.