11th International Congress
of Logic, Methodology
and Philosophy of Science
August 20-26, 1999
Cracow, Poland
Section 12: Philosophy of Cognitive Sciences and Artificial Intelligence

ARCHITECTURE-BASED CONCEPTIONS OF MIND

AARON SLOMAN

http://www.cs.bham.ac.uk/~axs/
A.Sloman@cs.bham.ac.uk

Ideas developed

in collaboration with

Steve Allen, Luc Beaudoin, Brian Logan, Catriona Kennedy, Ian Millington, Riccardo Poli, Ian Wright,

and others in the

COGNITION AND AFFECT PROJECT

SCHOOL OF COMPUTER SCIENCE
THE UNIVERSITY OF BIRMINGHAM

PROBLEM:
Do we understand what we mean by "consciousness" "emotion" "intelligence" "mind" etc .... ???

SOME INADEQUATE APPROACHES

1. Definitions in terms of behaviour and behavioural dispositions

These don't work because any collection of behaviours (and behavioural dispositions) can arise out of arbitrarily many different causal mechanisms.

2. Ostensive definitions based on "first person" experience.

These don't work, though they seduce many scientists and philosophers. Being able to recognize a subset of instances and non-instances does not require a full understanding of the general principles involved.

(Compare thinking you have a grasp of the concept of simultaneity because you have first-hand "direct" experience of simultaneity.)

SUGGESTIONS FOR MAKING PROGRESS:

(a) We need to see concepts of mind as "cluster concepts"

(b) We need to see them as "architecture-based" concepts

(c) The relevant architectures are *virtual machine* architectures implemented in but importantly different from *physical machines*.

EXAMPLE: "EMOTION"
Different definitions in psychology, philosophy, neuroscience, ethology ...

and many variants within each discipline

PARTIAL DIAGNOSIS:
Different theorists concentrate on different phenomena.
We need a theory that encompasses all of them.

REPHRASE:

1. What are the architectural requirements for various kinds of mental states and processes in humans and other animals?

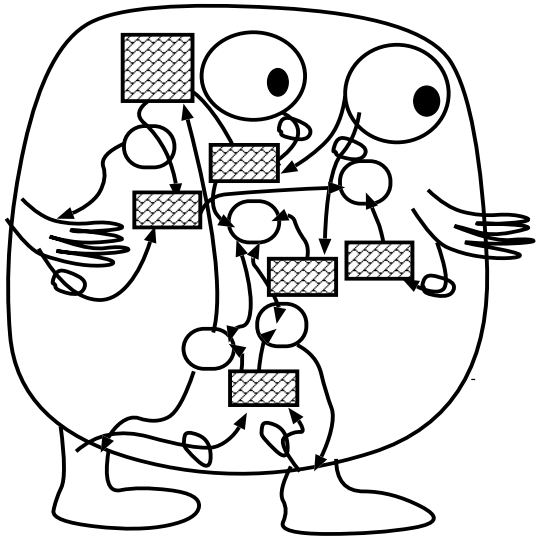
2. What sorts of states and processes can each architecture support?

Collect examples of many types of real (and theoretically possible) phenomena.
Try to build a theory which explains them all!
Subject to constraints from neuroscience, psychology, biological evolution, feasibility, tractability, etc.

ALLOW FOR VARIATION (different clusters of capabilities):

- **Across species,**
- **Within species,**
- **Within an individual during normal development**
- **After brain damage**
- **Across planets (grieving, infatuated, Martians?)**
- **Across the natural/artificial divide**

WHAT SORT OF ARCHITECTURE? COULD IT BE AN UNINTELLIGIBLE MESS?



YES, IN PRINCIPLE.

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

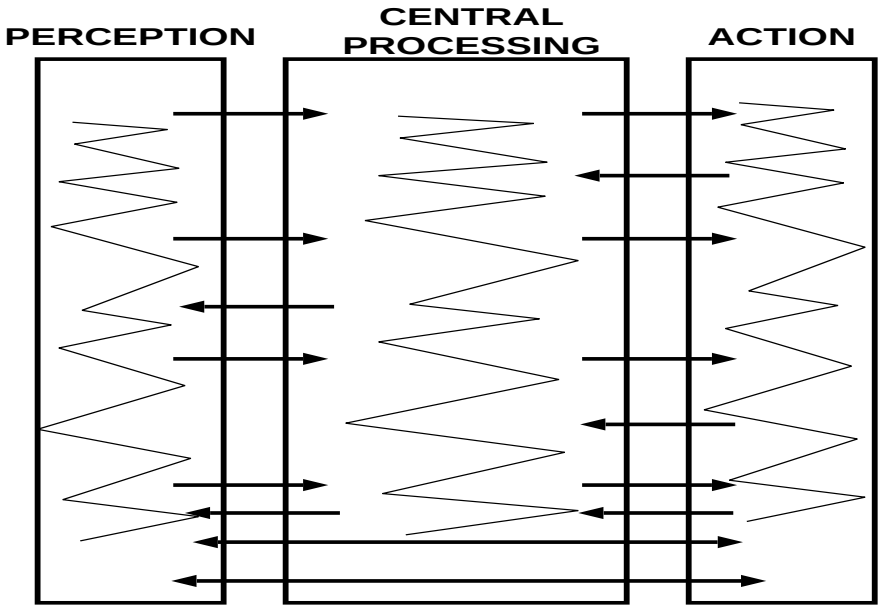
Problem 1: time required and variety of contexts required for a suitably general design to evolve.

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

TOWARDS A UNIFYING MODULAR THEORY OF BRAIN AND MIND: A BIRD'S EYE VIEW

One perspective:

THE "TRIPLE TOWER" MODEL



(many variants)
(Nilsson, Albus)

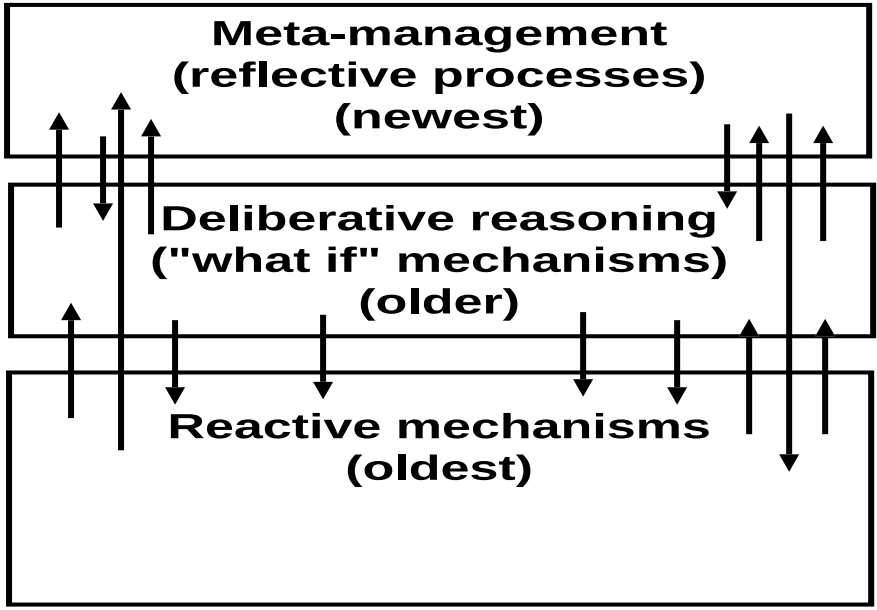
MODULAR **does not mean** RIGID **or** INNATE

Systems can be "nearly decomposable". Boundaries can change with learning and development.

ANOTHER COMMON ARCHITECTURAL PARTITION

(functional, evolutionary)

THE "TRIPLE LAYER" MODEL



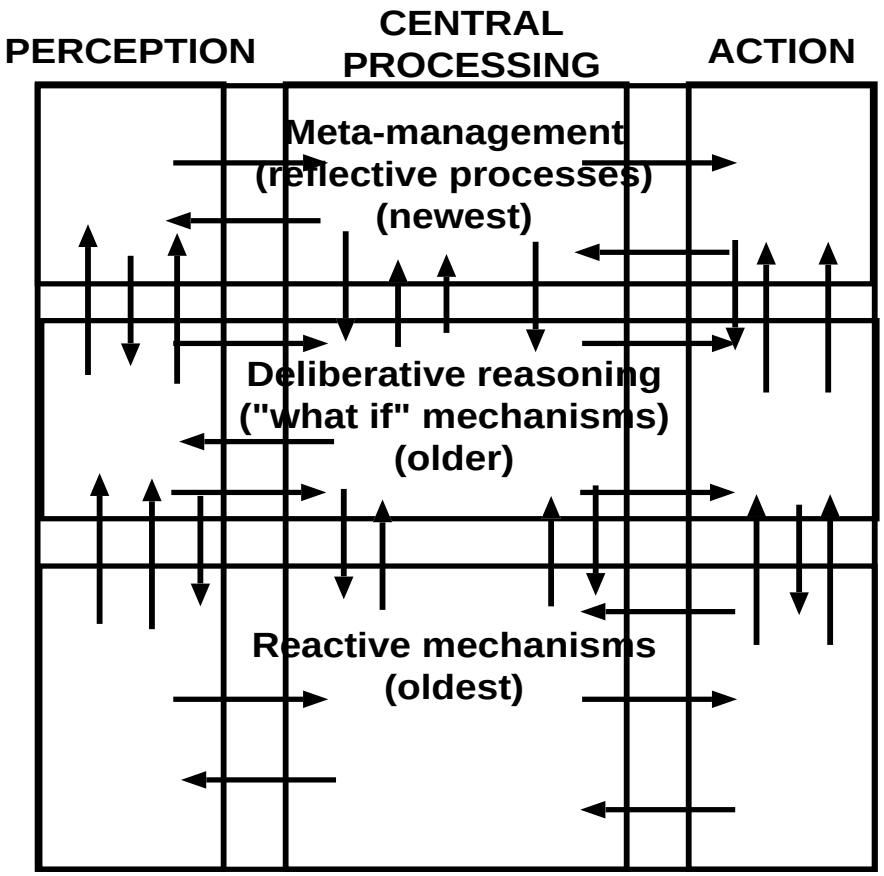
(many variants – for each layer)

Reactive systems can be highly parallel, very fast, and use analog circuits.

Deliberative mechanisms are inherently slow, serial, knowledge-based, resource limited.

COMBINING THE VIEWS: LAYERS + PILLARS = GRID

A grid of co-evolving sub-organisms, each contributing to the niches of the others.



SENSING AND ACTING CAN BE ARBITRARILY SOPHISTICATED

- **Don't treat sensors and motors as mere transducers.**
- **They can have sophisticated information processing architectures.**

  E.g. perception and action can be hierarchically organised with concurrent interacting sub-systems.

- **Perception goes far beyond segmenting, recognising, describing what is "out there". It includes:**
  - **providing information about *affordances* at different levels of abstraction. (Think of Gibson, not Marr),**
  - **directly triggering physiological reactions (e.g. posture control, sexual responses)**
  - **evaluating what is detected,**
  - **triggering new motivations**
  - **triggering "alarm" mechanisms**
  - **. . . . .**

AN EXTENSION OF GIBSON'S THEORY:

Different sub-systems use different affordances, and different ontologies. (Evidence from brain damage.)

They rely on processing by different virtual machines:

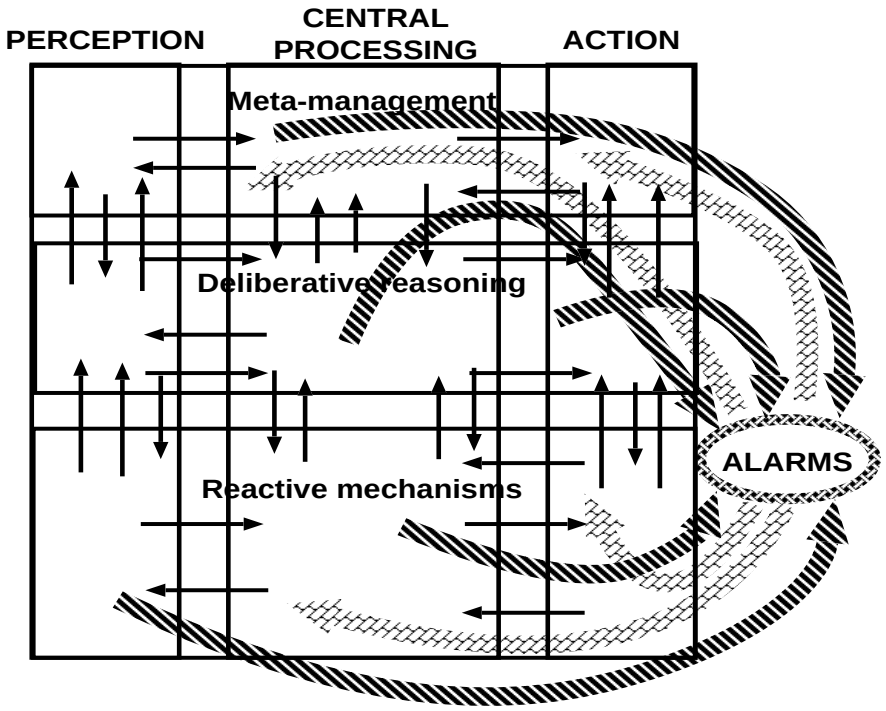
WITTGENSTEIN:

The substratum of an experience is mastery of a technique (mostly unconscious) (Compare Ryle)

As processing grows more sophisticated, so it can be come slower, to the point of danger.

FAST, POWERFUL,
"GLOBAL ALARM SYSTEM"
NEEDED

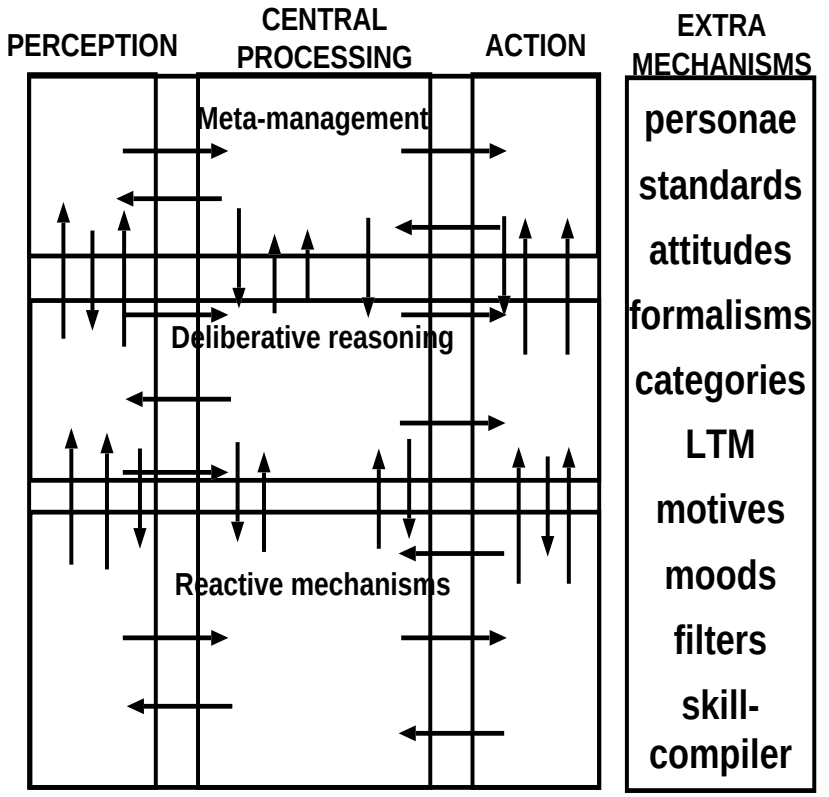
IT WILL INEVITABLY BE STUPID!



MANY VARIANTS POSSIBLE.

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

ADDITIONAL COMPONENTS
(No time to discuss)



MANY PROFOUND IMPLICATIONS

e.g. for kinds of development
kinds of perceptual processes
kinds of brain damage
kinds of emotions

VARIETIES OF MOTIVATIONAL SUB-MECHANISMS

MOTIVATION IS NOT JUST ONE THING

Motives or goals can short term, long term, permanent.

They can be triggered by physiology, by percepts, by deliberative processes, by metamanagement.

So there are many sorts of motive generators: MG

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 ?

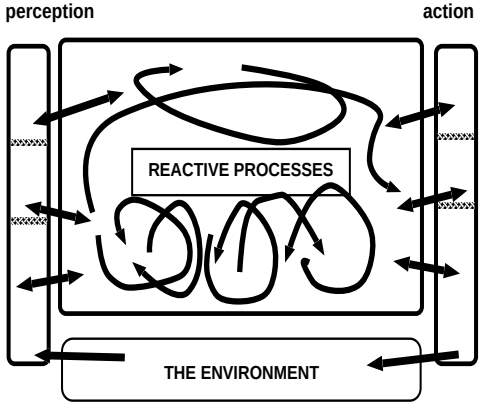
There are also EVALUATORS.

Current state can be evaluated as good, or bad, to be preserved or terminated. (Important for learning.)

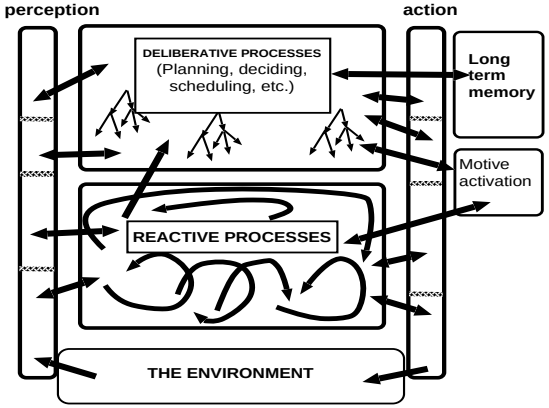
These evaluations can occur at different levels in the system, and in different subsystems, accounting for many different kinds of pleasures and pains. (Often confused with emotions.)

NOT ALL PARTS OF THE GRID ARE PRESENT IN ALL ANIMALS

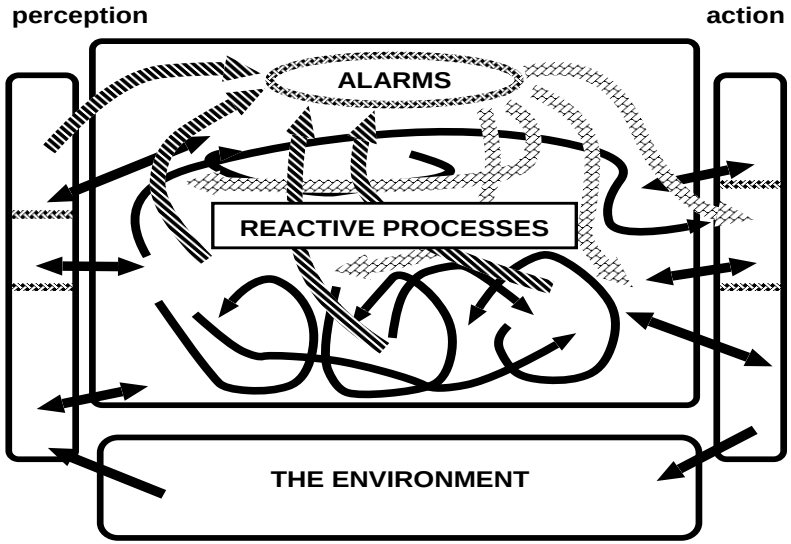
How to design an insect?



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



EMOTIVE INSECTS?

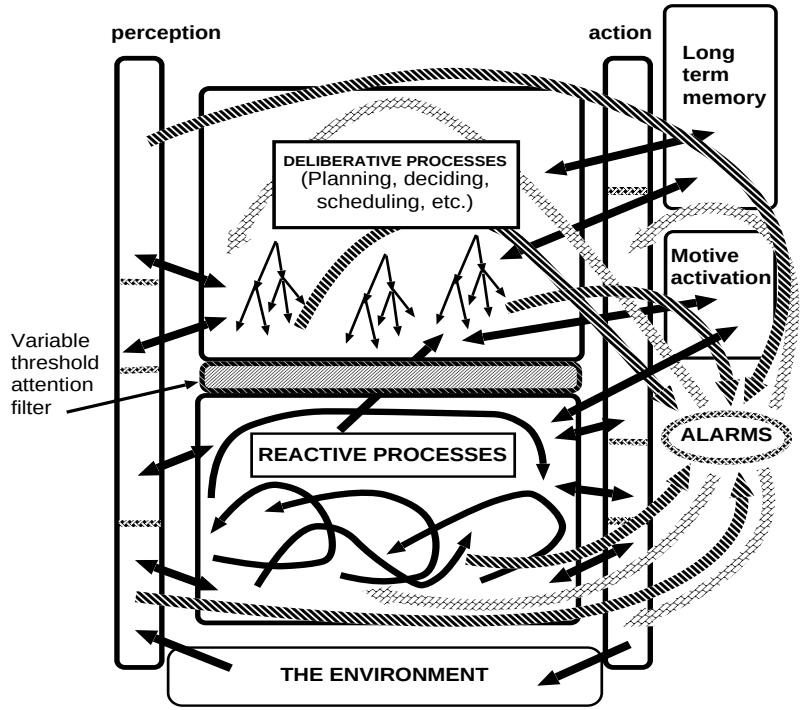


ALARM MECHANISM (GLOBAL INTERRUPT/OVERRIDE):

- **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

Damasio and Picard call certain states generated in reactive mechanisms via global alarm systems "Primary Emotions"

REACTIVE AND DELIBERATIVE LAYERS WITH ALARMS

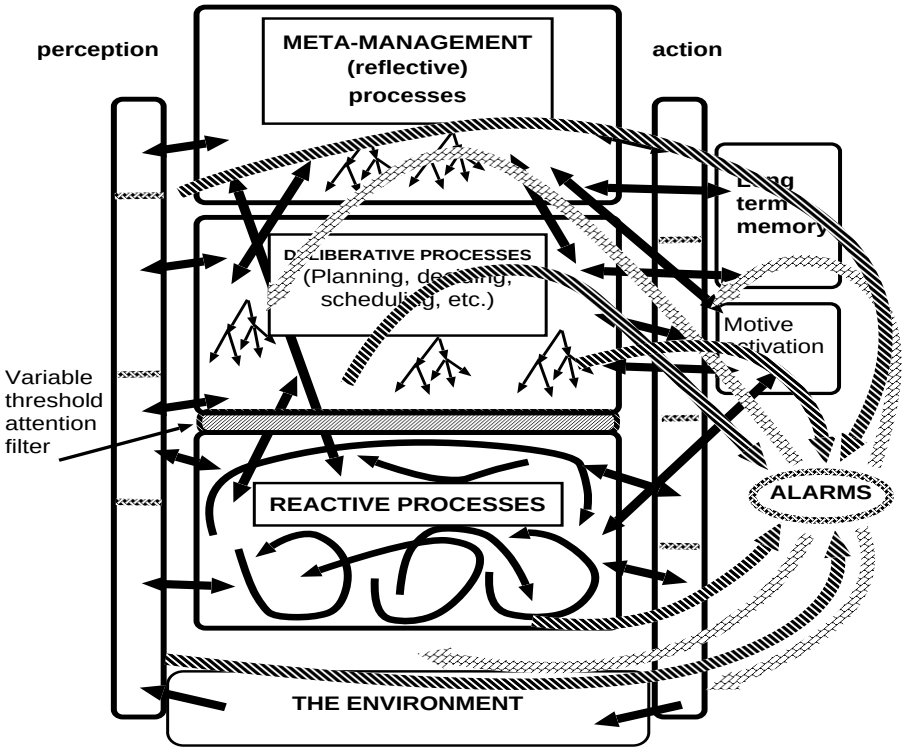


AN ALARM MECHANISM (Brain stem, limbic system?):
Allows rapid redirection of the whole system

- **Freezing, fleeing, arousal etc. as before**
- **Becoming apprehensive about anticipated danger**
- **Rapid redirection of deliberative processes.**
- **Relief at knowing danger has passed**
- **Specialised learnt responses: switching modes of thinking.**

Damasio & Picard:
cognitive processes trigger "secondary emotions".

METAMANAGEMENT WITH ALARMS



Tertiary emotions (previously called "perturbances") involve interruption and diversion of thought processes.

I.e. the metamanagement 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.

THESE LAYERS EXPLAIN

PRIMARY, SECONDARY, TERTIARY EMOTIONS

Different architectural layers support different sorts of emotions, and help us define

AN ARCHITECTURE-BASED ONTOLOGY OF MIND

Different animals will have different mental ontologies

Humans at different stages of development will have different mental ontologies

The REACTIVE **layer with** GLOBAL ALARMS **supports "primary" emotions:**

- **being startled**
- **being disgusted by horrible sights and smells**
- **being terrified by large fast-approaching objects?**
- **sexual arousal? Aesthetic arousal ?**

etc. etc.

The DELIBERATIVE **layer enables "secondary" emotions (cognitively based):**

- **being anxious about possible futures**
- **being frustrated by failure**
- **excitement at anticipated success**
- **being relieved at avoiding danger**
- **being relieved or pleasantly surprised by success**

etc. etc.

THE THIRD LAYER
enables
SELF-MONITORING,
SELF-EVALUATION
and
SELF-CONTROL

AND THEREFORE ALSO LOSS OF CONTROL (PERTURBANCE)
(and qualia!)

This makes possible "tertiary" emotions, through having and losing control of thoughts and attention:

- **Feeling overwhelmed with shame**
- **Feeling humiliated**
- **Aspects of grief, anger, excited anticipation, pride,**
- **Being infatuated, besotted**

and many more *typically HUMAN* **emotions.**

NOTES:

1. Different aspects of love, hate, jealousy, pride, ambition, embarrassment, grief, infatuation can be found in all three categories.

2. Remember that these are not STATIC **states but** DEVELOPING **processes, with very varied aetiology.**

SOCIALLY IMPORTANT
HUMAN EMOTIONS
INVOLVE RICH CONCEPTS
AND KNOWLEDGE
and
RICH CONTROL MECHANISMS
(architectures)

- **Our everyday attributions of emotions, moods, attitudes, desires, and other affective states implicitly presuppose that people are information processors.**
- **To long for something you need to know of its existence, its remoteness, and the possibility of being together again.**
- **Besides these *semantic* information states, longing also involves *control* states.**

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.

- **Physiological processes (outside the brain) may or may not be involved. Their importance is normally over-stressed by experimental psychologists under the malign influence of the James-Lange theory of emotions. (Contrast Oatley, and poets.)**

CONCLUSION: THE SCIENCE

- **Much of this is conjectural – many details still have to be filled in and consequences developed (both of which can come partly from building working models, partly from multi-disciplinary empirical investigations).**
- **An architecture-based ontology can bring some order into the morass of studies of affect (e.g. myriad definitions of "emotion").**

  COMPARE THE RELATION BETWEEN THE PERIODIC TABLE OF ELEMENTS AND THE ARCHITECTURE OF MATTER.

- **This can lead to a better approach to comparative psychology, developmental psychology (the architecture develops after birth), and effects of brain damage and disease.**
- **It will provide a conceptual framework for discussing which kinds of emotions can arise in software agents that lack the reactive mechanisms required for controlling a physical body.**

CONCLUSION: ENGINEERING

Designers need to understand these issues:

(a) if they want to model human affective processes,

(b) if they wish to design systems which engage fruitfully with human affective processes,

(c) if they wish to produce teaching/training packages for would-be counsellors, psychotherapists, psychologists.

(d) and maybe even for convincing synthetic characters in computer entertainments?

COGNITION AND AFFECT PROJECT PAPERS:

ftp://ftp.cs.bham.ac.uk/

pub/groups/cog_affect/0-INDEX.htm

AND OUR TOOLS:

ftp://ftp.cs.bham.ac.uk/

pub/dist/poplog/freepoplog.htm

(Including the SIM_AGENT toolkit)