OUP/Prospect Debate, London 17 Nov 1998

ARE BRAINS COMPUTERS?

AARON SLOMAN

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

SCHOOL OF COMPUTER SCIENCE
THE UNIVERSITY OF BIRMINGHAM

Obviously brains are not artefacts and obviously they are not built like contemporary computers.

So, to avoid triviality, let's ask:

WHAT KIND OF MACHINE CAN HAVE MENTAL STATES?

What are the architectural requirements for human-like mental states and processes?

RELATED THEMES (No time for full discussion):

Emergence of non-reducible layers in reality.
The role of virtual machines in explaining mental processes.
How different architectures support different kinds of minds.
with different sorts of emotions, thoughts, etc.
The evolution of mind.
How evolution itself involves an abstract emergent virtual machine with multiple interacting feedback loops.
Minimal requirements for physical underpinning.

ASK THE GHOST OF GILBERT RYLE

What sort of ghost in the machine can think?

WHAT SORT OF MACHINE?

AN INFORMATION PROCESSING MACHINE!

Gilbert Ryle's 1949 book THE CONCEPT OF MIND **had many important and relevant ideas about internal information processing, e.g. in his chapter on imagination.**
But he was wrongly interpreted as a behaviourist.

He was (unwittingly) discussing information processors.

THREE SORTS OF MACHINES
(These are not mutually exclusive.)

• ENERGY MANIPULATORS:

machines which transform, store, transmit or use energy. E.g. steam engines, electric heaters, electricity generators, candles, gears, rockets.

• MATTER MANIPULATORS:

machines which transform matter, at the chemical level or on a larger scale, creating, disassembling or recombining structures of various sorts, e.g. digesting food, assembling cars.

• INFORMATION MANIPULATORS:

machines which acquire, analyse, transform, store, manipulate, interpret, transmit and use information.

COMPARE

WHEN A HUGE ASTEROID HITS THE EARTH:

it causes devastating movements of MATTER **and** ENERGY **all over the earth.**

WHEN DIANA DIED:

it was INFORMATION **that flowed around the planet causing much grief, consternation, curiosity, re-scheduling of television programs, and production of huge amounts of printed material.**

INFORMATION CHANGES CAN HAVE CAUSAL POWERS

EVEN IN A VIRTUAL MACHINE

WHAT'S AN INFORMATION PROCESSING MACHINE?

AN INFORMATION PROCESSOR WILL HAVE SOME COMBINATION OF THE FOLLOWING FEATURES

- IT INCLUDES STRUCTURES WHICH CHANGE OVER TIME. **(Many possible formats: bit patterns, continuously changing physical values, lists of symbols, sentences compatible with a particular grammar, 2-D image structures, trees, networks, etc.)**
- IT CAN CREATE NEW STRUCTURES BASED ON (DERIVED FROM) OLD ONES. **(E.g. rule-driven copying, concatenating, abbreviating, deriving.)**
- IT CAN MODIFY OLD STRUCTURES **(e.g. replacing a part with something different, of similar, greater, or lesser complexity, adding links between structures, changing a continuous variable)**
- IT CAN "INTERPRET" SOME OF ITS STRUCTURES AS INSTRUCTIONS FOR MANIPULATING STRUCTURES **(e.g. deciding which structures to operate on, selecting which operations to perform, deciding in which locations to store results of the operations).**
- IT CAN DERIVED NEW STRUCTURES FROM SENSORY INPUT. **(E.g. receiving email via an internet connection, or image structures from a TV camera.)**
- IT CAN TRANSMIT STRUCTURES **to physical machines or to other information processing systems via output transducers. (E.g. sending control signals to motors, sending information structures to other information processors.)**
- SEVERAL OF THE ABOVE MAY HAPPEN CONCURRENTLY AND INFLUENCE ONE ANOTHER**, within the same system.**

LOW LEVEL REQUIREMENTS FOR INFORMATION PROCESSORS

- **Rich variability of state (many switchable components)**
  If N is large 2^N can be astronomical
- **High connectivity (physical or virtual)**
  Compare connectivity of neurons and random access in RAM
- **Causal connections between state changes (IF ... THEN ...)**
- **Internal & external control of internal & external behaviour**
- **Self modification (sometimes self construction???)**
- **Compactness (for mobile organisms and robots)**
- **Low energy consumption**
- **Long term stability (e.g. resistance to thermal buffeting)**
- **Special purpose processors (e.g. image processors)**

NOT NECESSARILY DISCRETE and SERIAL

Various kinds of hybrids are possible.

However discrete multi-stable elements have advantages for long term memory. (Bi-stable elements are just a special case).

Moreover, deliberative mechanisms may be *inherently* serial in their high level functionality. (EXPLAINED LATER)

New kinds are being explored: Optical computers, DNA computers, Quantum computers,...

EVOLUTION BEAT US BY MILLIONS OF YEARS:
strength/weight ratios, fuel stores, efficient energy conversion in motors, various forms of propulsion (land, sea and air) AND
powerful information processing mechanisms.

BRAINS SUPPORT CONSCIOUSNESS
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.

Not just ONE thing.

THEY 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.

I.E. "CONSCIOUSNESS" IS A VERY VAGUE "CLUSTER CONCEPT".

(Like "emotion")

People think they know what IT is from experience. Before Einstein people thought they knew what simultaneity was. We can unintentionally fool ourselves.

WHAT ARE THE CAPABILITIES?

There are several types, of varying evolutionary age, including:

- REACTIVE (VERY OLD, VERY WIDESPREAD)

  Reactive capabilities, including innate reflexes and learnt or trained responses.
  There are many kinds, of varying degrees of sophistication, e.g. some with learning, some with goal-processing, some using explicit stored plans.

- DELIBERATIVE (NEWER, NOT SO WIDESPREAD)

  These require special kinds of content addressable associative memory, special kinds of temporary workspace, special abilities to construct nested structures.

- REFLECTIVE, META-MANAGEMENT (NEWEST, RAREST?)

  The ability to monitor, evaluate and to control or redirect internal processes, e.g. redirecting attention, changing thinking strategy or topic.

- GLOBAL INTERRUPTS (ALARMS)

  Initially parts of reactive systems, but become more elaborate as more architectural layers are added.

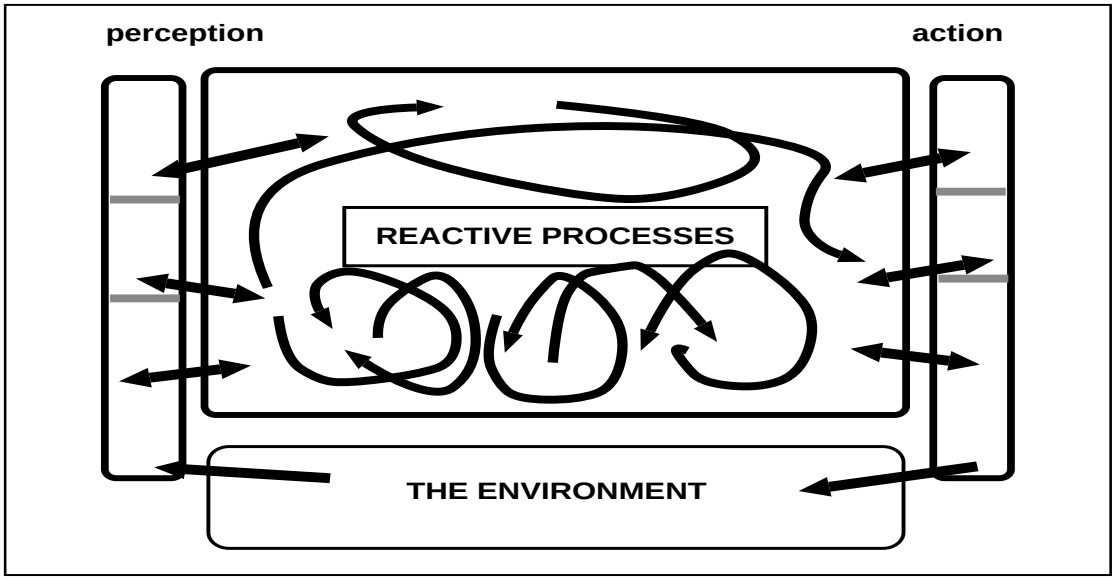
- PERCEPTUAL AND ACTION MECHANISMS
  (Operating at different levels of abstraction)

The different layers operate concurrently, and are sometimes mutually supportive, sometimes not.

DIFFERENT COMBINATIONS OF CAPABILITIES CAN BE IMPLEMENTED IN DIFFERENT SORTS OF ARCHITECTURES.

NB THESE ARE **NOT** DIFFERENCES OF DEGREE

REACTIVE AGENTS
How to design an insect?



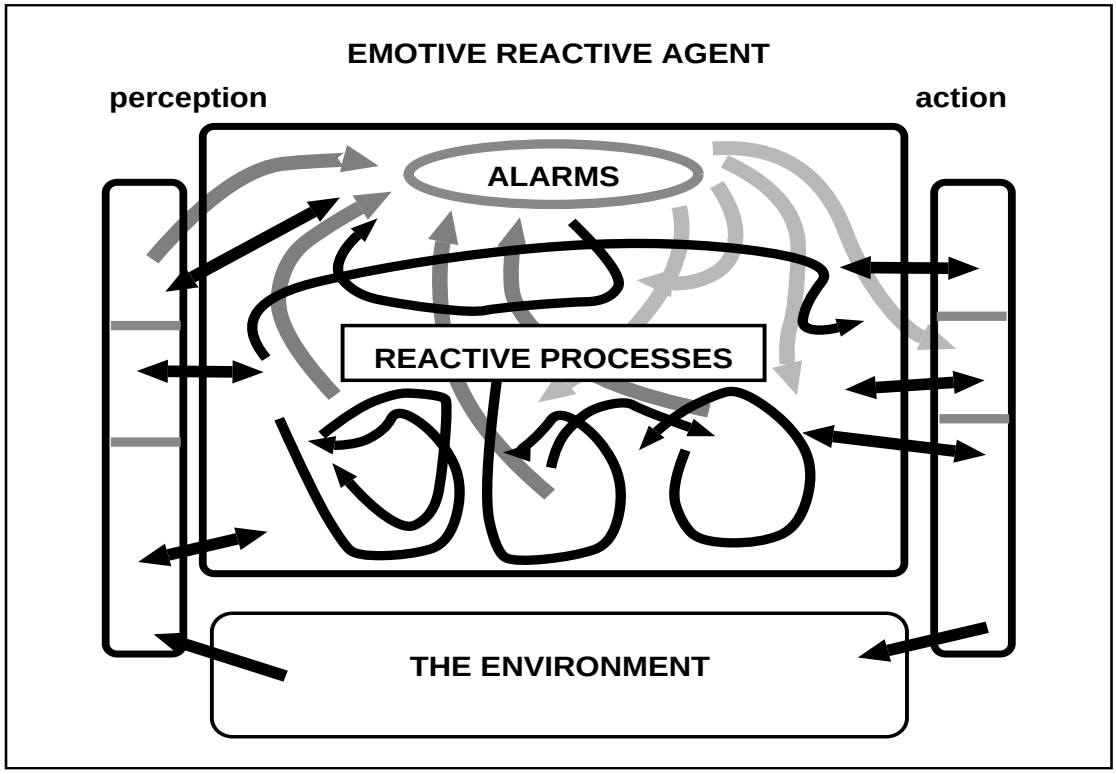
IN A REACTIVE AGENT:

- **Mechanisms and space are dedicated to specific tasks**
- **No construction of new plans or structural descriptions**
- **No explicit evaluation of alternative structures**
- **Conflicts handled by vector addition, simple rules or winner-takes-all nets.**
- **Parallelism and dedicated hardware give speed**
- **Many processes may be analog (continuous)**
- **Agents cope using only genetically determined behaviours**
- **Some learning is possible: e.g. tunable control loops, change of weights by reinforcement learning**
- **Cannot cope if environment requires new plan structures.**
- **Compensate by having large numbers of expendable agents?**

There are different classes of reactive architectures.

Some use several processing layers: e.g. high order control loops.
Some manipulate internal state.

EMOTIVE REACTIVE AGENTS

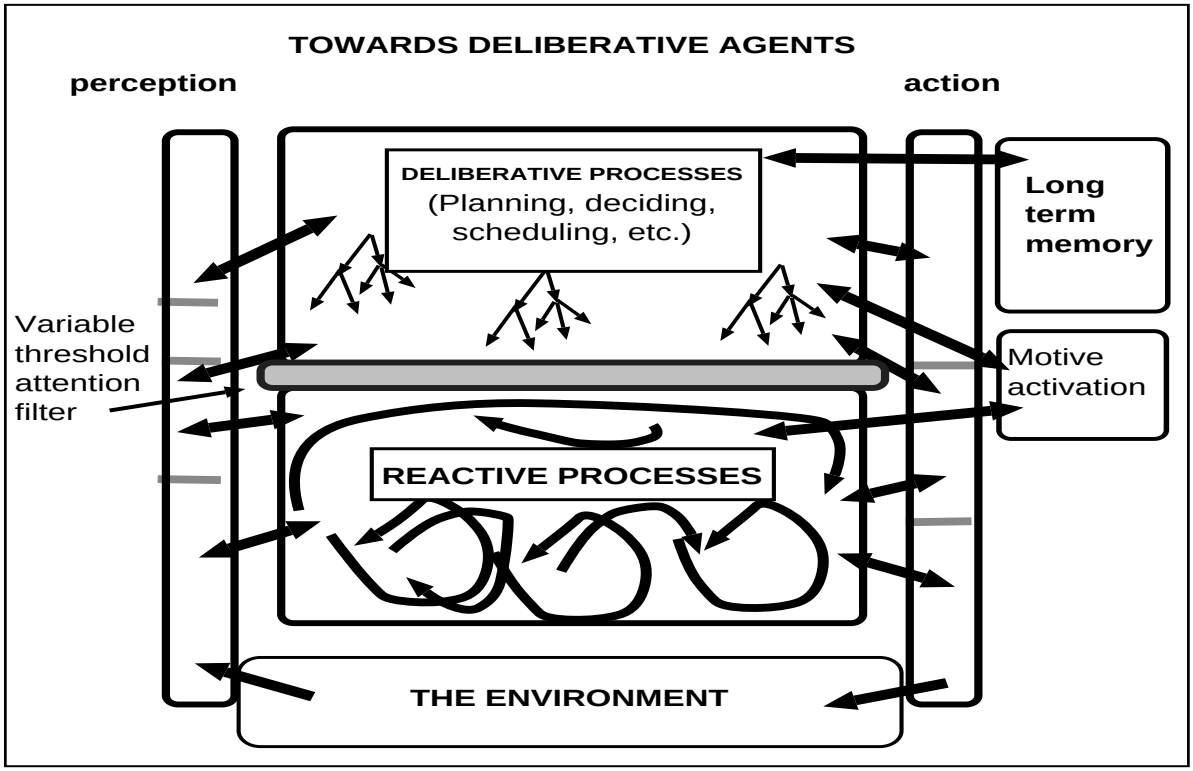


AN ALARM MECHANISM (GLOBAL INTERRUPT/OVERRIDE):

- **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 these "Primary Emotions"

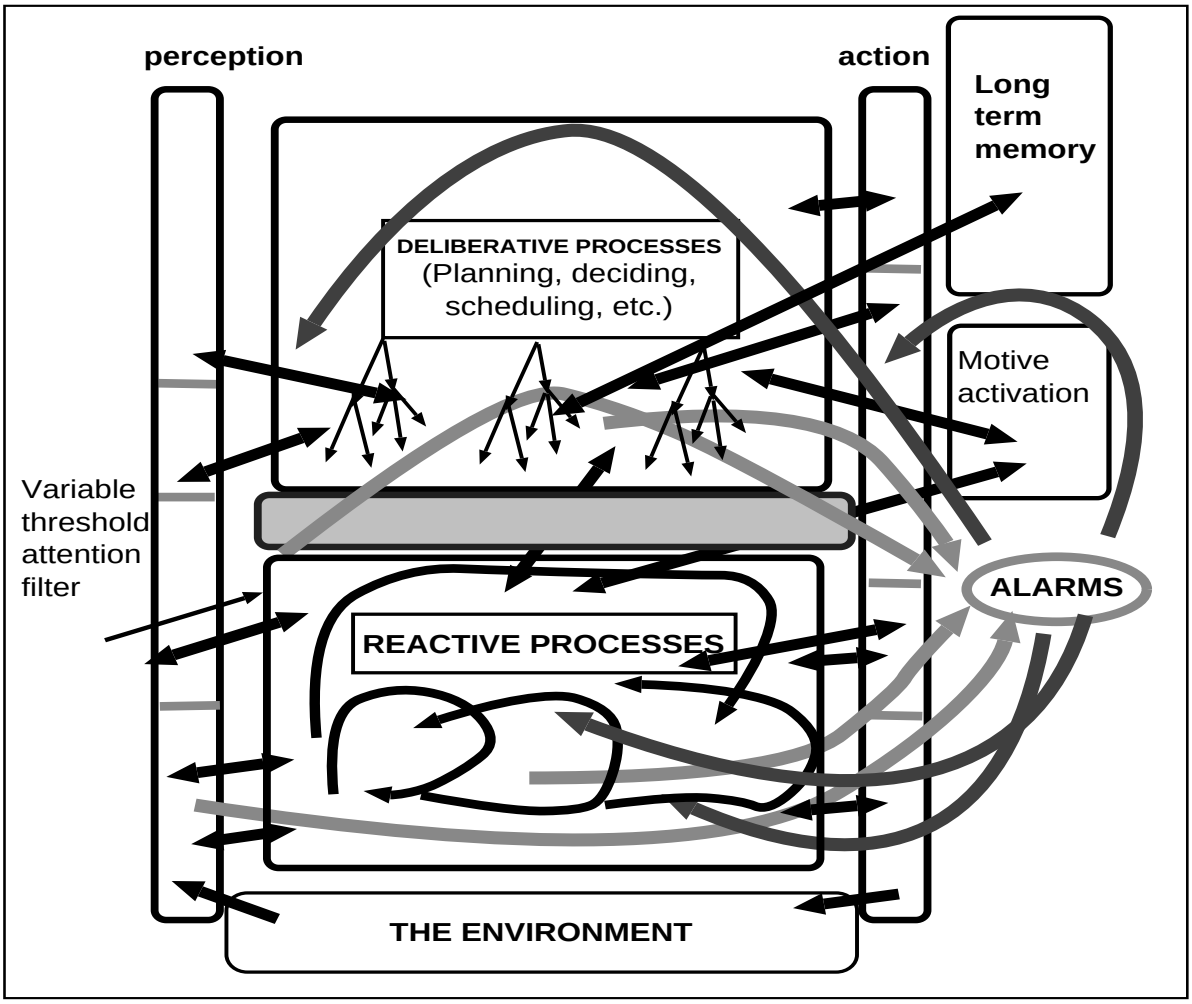
ADD A DELIBERATIVE LAYER



IN A DELIBERATIVE MECHANISM:

- **Motives are explicitly generated and plans created**
- **New options are constructed and evaluated**
- **Mechanisms and space are reused serially**
- **Learnt skills can be transferred to the reactive layer**
- **Sensory and action mechanisms may produce or accept more abstract descriptions (hence more layers)**
- **Parallelism is much reduced (for various reasons):**
  - LEARNING REQUIRES LIMITED COMPLEXITY
  - SERIAL ACCESS TO (PARALLEL) 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.**

REACTIVE AND DELIBERATIVE LAYERS WITH ALARMS

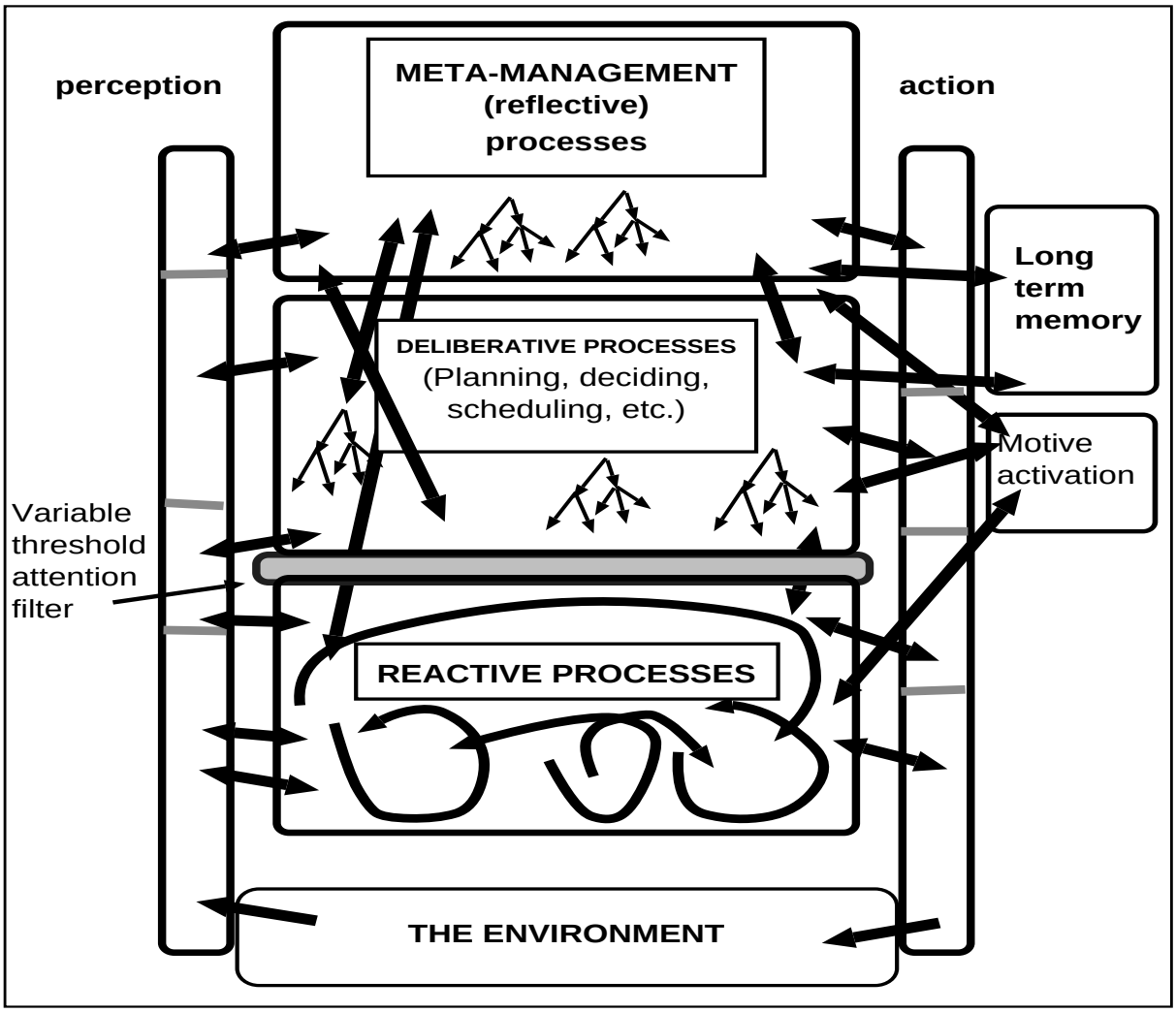


AN ALARM MECHANISM (Brain stem, limbic system?):
Allows rapid redirection of the whole system

- **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".

REFLECTIVE AGENTS



META-MANAGEMENT ALLOWS

- **Self monitoring (of many internal processes)**
- **Self evaluation**
- **Self modification (self-control)**

NB: ALL MAY BE IMPERFECT

- **You don't have full access to your inner states and processes**
  IT'S JUST ANOTHER TYPE OF PERCEPTION!
- **Your self-evaluations may be ill-judged**
- **Your control may be partial (why?)**

SELF-MONITORING (META-MANAGEMENT)

Deliberative mechanisms with evolutionarily determined strategies may be too rigid.

Internal monitoring and evaluation mechanisms may help:

- **Improve the allocation of scarce deliberative resources e.g. detecting "busy" states and raising interrupt threshold,**
- **Record events, problems, decisions taken by the deliberative mechanism,**
- **Detect management patterns, such as that certain deliberative strategies work well only in certain conditions,**
- **Allow exploration of new internal strategies, concepts, evaluation procedures, allowing discovery of new features, generalisations, categorisations,**
- **Allow diagnosis of injuries, illness and other problems by describing internal symptoms to experts,**
- **Evaluate high level strategies, relative to high level long term generic objectives, or standards.**
- **Communicate more effectively with others, e.g. by using viewpoint-centred appearances to help direct attention, using drawings to communicate how things look.**

HAVING ACCESS TO CONTENTS OF INTERMEDIATE PERCEPTUAL BUFFERS WILL CAUSE SOME ROBOTS TO DISCOVER QUALIA!

THESE LAYERS EXPLAIN

PRIMARY, SECONDARY, TERTIARY EMOTIONS

Different architectural layers support different sorts of emotions:

The REACTIVE **layer with** GLOBAL ALARMS **supports:**

- **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:**

- **being anxious or apprehensive 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 SELF MONITORING META-MANAGEMENT **layer, explains:**

- **having and losing control of thoughts and attention:**

Feeling ashamed of oneself
Feeling humiliated
Aspects of grief, anger, excited anticipation, pride,
Being infatuated, besotted
and many more typically HUMAN emotions.

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

FEARERS AND DOUBTERS
regarding computational models

TWO KINDS OF OBJECTORS

1. Fearers: the longed for gap

Some people simply don't *like* the idea of machines (as they construe them) ever being so much like us, because of

- **Fear of machines taking control**
- **Concerns about loss of human dignity (ontological neurosis, theological worries)**

2. Doubters: the perceived gap

Doubters see the limitations of existing computer-based machines and software systems and cannot imagine any ways of overcoming them.

They accept over-simple views of computers: e.g. as able to do only what they have been programmed to do.

Partial answer to doubters:

1. The full potential of computers, especially networks of asynchronous cooperating computers, is something we barely understand.

2. We are steadily learning how to make them more and more sophisticated, but we have a long way to go.

3. Doubters may, in the long run, turn out to be correct, if, for instance, the functioning of animal brains turns out to require some kind of mechanism we have not yet dreamed of.

FEARERS and DOUBTERS COMPUTE AND RECONSIDER

1. TO FEARERS:

- **Your ideas on human dignity may focus on the wrong aspects of dignity: it's what we are that matters, not what we are made of or where we came from.**
- **Machines could never be more horrible in their behaviour than the worst humans.**
- **If your worries are theological, try to abandon your theology. It's probably false anyway. (Get a life, instead of worrying about the afterlife.)**

2. TO DOUBTERS:

- **Stretch your ideas about mechanisms, especially information processing mechanisms.**
- **Analyse your concepts of mind.**
- **Then think again.**

THE DOUBTERS MAY TURN OUT TO BE RIGHT.

But it's an EMPIRICAL question.

Armchair arguments don't work.

Except for wishful thinkers.

We need to study:

- **brains (of all kinds)**
- **computations (of all kinds).**

HOW TO MAKE PROGRESS

IF we discover that brains use powerful mechanisms we currently don't understand,

THEN we'll find out how they work,
and incorporate those mechanisms into new computers
E.g. DNA computers, quantum computers

The key idea about computers

- **is not the technology (which constantly changes), BUT**
- **the nature of what they *do*.**

BOTH **are, above all, information processors.**

IN BOTH **capabilities depend on information they have.**

We don't yet know what information or architectures they require for many tasks we find easy, e.g. seeing.

They are still inferior in many respects (but by no means all).

Some gaps are steadily diminishing.

There's even work on emotions in synthetic agents.

The kinds of information they can handle, and they ways they handle information are constantly being extended.

IN PARTICULAR:

- **The causal interactions *within* information processing virtual machines are constantly being extended.**

- **Having an experience leading to a thought producing a desire, producing further thoughts, producing emotions is just a chain of causes and effects within an information processing virtual machine.**

CONCLUSION

If you don't believe in magic, why believe the gap will never be closed?

I DON'T BELIEVE IN MAGIC.

BRAINS ARE INFORMATION PROCESSORS

I CONCLUDE

BRAINS ARE COMPUTERS,

THOUGH AT PRESENT

WE DON'T YET KNOW

WHAT KINDS OF COMPUTERS.