#### Digital Biota 2 http://www.cyberbiology.org

## WHAT SORTS OF BRAINS CAN SUPPORT WHAT SORTS OF MINDS? Aaron Sloman http://www.cs.bham.ac.uk/~axs/ A.Sloman@cs.bham.ac.uk School of Computer Science The University of Birmingham

**Including ideas from:** 

Brian Logan, Riccardo Poli, Luc Beaudoin, Ian Wright, And many others at Birmingham and elsewhere.

#### THE COGNITION AND AFFECT PROJECT DIRECTORY

ftp://ftp.cs.bham.ac.uk/pub/groups/cog\_affect/0-INDEX.html

This talk has many related themes: emergence of non-reducible layers in reality, the role of virtual machines in explaining intelligence, the evolution of mind, how different architectures support different kinds of minds, how evolution itself involves an abstract emergent virtual machines with multiple interacting feedback loops.

### **TWO KINDS OF RESEARCHERS**

Discussion of the relationship between mind and body often gets bogged down because of disagreements between two kinds of thinkers:

TYPE 1: search for interesting abstract levels at which things are the same, and general principles at these levels.

TYPE 2: study the interesting details to be found in special cases.

For Type1 researchers it is obvious that:

• artificial life and biological life can be construed as essentially the same sort of thing,

• there is no essential difference between intelligent software systems, intelligent robots and intelligent animals

• simulated environments and physical environments are not importantly different.

For Type2 thinkers it is a serious intellectual sin to lump together such obviously diverse phenomena, ignoring all the important details in which they differ, and which require different principles and behavioural laws for their description and explanation. Is any kind of synthesis of these two views possible?

Not where the clash is driven by emotional and motivational differences or theological concerns.

From a Type 1 viewpoint, these two principles can provide a synthesis:

(A) THE PRINCIPLE OF MULTIPLE LEVELS OF REALITY.

(B) THE PRINCIPLE OF MULTIPLE REALISATIONS OF INTELLIGENCE.

The point is that things that differ at a low level (e.g. in the underlying mechanisms used and other implementation details) may have the same high level features, and even very similar virtual machines.

E.g. from this viewpoint the propellers and the wings on an aeroplane are essentially doing the same job as the wings on a bird, i.e. providing thrust and lift in accordance with the same principles of aerodynamics and the same tradeoffs between stability and manoeuvreability.

### **MULTIPLE LEVELS OF REALITY**



There are many kinds of causally efficacious "emergent" virtual machines, operating at different scales and different levels of abstraction.

- Is there a well-defined "bottom level"? We don't know whether physics has a well defined bottom level. It seems to have several levels.
- Causality operates at all levels.
- Causality can be "circular". (E.g. physical processes in a computer cause events in a computational virtual machine, and non-physical events in the virtual machine, e.g. a queen capturing a bishop, causes physical events.)

• The biosphere is a richly interacting collection of myriad virtual and physical machines running concurrently.

• Levels need not form a rigid hierarchy.

**Terminology:** 

PHILOSOPHERS: talk about "supervenience"

**ENGINEERS:** refer to the same thing as "implementation"

### (a) Multiple levels of reality.

Besides physical objects, events, processes, there are also such things as species, genes, poverty, crime, economic recessions, brains, beliefs, experiences, emotions, etc.

These things are REAL and have CAUSAL POWERS, even if their existence depends on a physical substratum.

Our world contains machines of many types, at many levels of physical scale, at many levels of abstraction, with many different kinds of causal powers, and many different kinds of laws. This is one reason why we have many sciences.

#### **EXAMPLES OF NON-PHYSICAL CAUSATION:**

**1.** Ignorance can cause poverty. Poverty can cause crime and ignorance. Crime can cause death to people and damage to property.

2. Bugs in software can cause control instructions to be inappropriate. Inappropriate control instructions can cause a rocket to crash.

**3.** Noticing something can cause a person to have a new motive. Having a new motive can lead to reconsideration of old ones. Reconsideration can lead to revised decisions and new actions.

All of these depend on underlying processes in physical mechanisms. But to say that they are "nothing but" physical phenomena is to fall into the "nothing buttery" fallacy, ignoring the possibility of different kinds of co-existing, mutually supportive, reality.

#### WHAT ELSE IS THERE? STRUCTURE, INFORMATION, ...

If you lose a diskette containing the draft of a book you have worked on for ten years, and there is no other copy, you will not react kindly to someone who says it is "nothing but a collection of atoms and molecules, nothing but an inexpensive combination of plastic and metal oxide, which can easily be replaced.

## (b) Multiple realisations of intelligence.

### A very general and abstract concept of intelligence:

POSSESSION OF SOME ABILITY TO ACQUIRE, MANIPULATE

#### AND USE INFORMATION

In that sense even an amoeba has a kind of intelligence and so do very simple robots and many software systems, including various kinds of software "agents". Even a thermostat, or an operating system, has a limited type of intelligence in that sense.

Arguing about whether they are "really" intelligent is pointless. As pointless as arguing about whether they are "really" alive.

#### "Intelligence" (like "life") is a CLUSTER CONCEPT.

It refers to a complex cluster of capabilities, different combinations of which may be present in different intelligent systems. There's no well defined subset or disjunction of subsets constituting necessary and sufficient conditions for intelligence. Many of the capabilities themselves involve cluster concepts (e.g. the ability to see, to understand language, to remember, to solve problems, to learn....)

For related reasons we should not expect to find any well-defined *metric* for intelligence. At best we can expect different dimensions each with a partial ordering. E.g. partial orderings of:

- ability to see
- ability to detect things by smell
- ability to learn
- ability to plan
- ability to plan under pressure of time constraints
- fine control of intricate movements
- ability to communicate
- ability to notice "short cuts" in acting or thinking (productive laziness)

### UNDERSTANDING INTELLIGENCE REQUIRES US TO UNDERSTAND:

- 1. What sorts of "intelligences" there are,
- 2. How they differ,
- 3. What the implications of those differences are,
- 4. What sorts of designs can support which kinds,
- 5. How those designs can be implemented
- 6. What difference is made by different implementations of the same design
- 7. Which trade-offs drove the evolution of different kinds
- 8. Which mechanisms made that sort of evolution possible.

MYRIAD VARIETIES OF "NATURAL INTELLIGENCE" All use information, for different purposes and in different ways, requiring a host of different sorts of information processing architectures, sensing systems, processing mechanisms, processing strategies, forms of representation, divisions of labour between parts of one organism or within a colony or hive, etc.

VARIETIES OF "ENGINEERED INTELLIGENCES"

Yet more types of intelligence are being developed using computers, e.g. designing them, evolving them, letting them bootstrap themselves by learning, etc.

In part our task is to understand the variety of types of information processing *architectures* that are possible in principle, and to understand which are actually used in nature, and why, and which types are useful in new practical applications, and why.

Perhaps we can learn about possible extra-terrestial intelligences.

### **THREE TYPES OF MACHINES**

It is not always noticed that there are profound differences between three sorts of machines:

• ENERGY MANIPULATORS: machines which transform, store, transmit or use energy. E.g. steam engines, electric heaters, candles.

• MATTER MANIPULATORS: machines which transform matter, at the chemical level or on a larger scale, creating, disassembling or recombining structures of various sorts.

• INFORMATION MANIPULATORS: machines which acquire,

transform, store, manipulate, interpret and use information.

It appears that there were no information processing machines until the development of organisms.

Until recently only natural machines could process information, as opposed to transforming energy and matter (e.g. volcanoes, tornadoes, fires, stars).

Now computer-based machines of many kinds can do it, though not yet with the generality, flexibility and self-awareness of a human information processor.

### 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.

### WHAT IS EMERGENCE?

Like poverty, crime, thrashing in an operating system, capture of a piece in a chess virtual machine, the states and processes in an information processing architecture are not *physical*. "Emergent mechanisms" process information structures, like sentences, logical formulae, databases, visual arrays, list structures, information distributed in a neural net.

Emergence is not an anti-scientific metaphysical notion. Rather it refers to situations where there are (at least) two ontologies A (e.g. social or mental or computational events and processes) and B (e.g. physical or neural or digital electronic events and processes), both involving objects, events, processes, causal and other relations, etc.

#### **Ontology A is emergent on Ontology B where (roughly):**

- (a) Certain phenomena of type B SUFFICE FOR THE PRODUCTION OF phenomena of type A, in such a way that
- (b) Ontology A depends on but cannot be DEFINED in terms of Ontology B
- (c) Laws of A cannot be DERIVED from those of B. (This follows from (a)).

EXAMPLE: Chess playing virtual machines may be emergent on digital electronic machines.

MULTIPLE REALISABILITY: Sometimes one ontology can be based on (or implemented in) several different sorts of lower level ontologies.

Like the same word-processor running on different types of hardware, or a chess playing virtual machine running on different sorts of computer technology, and on different intermediate virtual machines (e.g. virtual machines for different programming languages: Prolog, C++, Java, etc.)

### **"SUPERVENIENCE" AND "IMPLEMENTATION"**

Precisely how to describe the different levels and their relationships can be very difficult.

Philosophers usually ignore cases like implementation of virtual machines in computers, and consequently produce incorrect theories of supervenience.

E.g. assuming that a regular correlation or processes or one to one mapping of structures is required between phenomena in the supervenient ontology and phenomena in the implementation ontology is incorrect.

By contrast, there are more subtle requirements that engineers learn: e.g. variability at the lower level must be able to support variability at the higher level except where economies are possible at the low because not all the theoretically possible high level variability will be encountered in practice. E.g. "sparse arrays" implement huge virtual machine structures in much "smaller" physical structures.

OFTEN NATURAL AND ARTIFICIAL SYSTEMS HAVE DEEP SIMILARITIES DESPITE SUPERFICIAL DIFFERENCES. A BIRD'S WING COMBINES THE FUNCTIONS OF AN AEROPLANE'S WING (LIFT) AND ITS PROPELLER (PROPULSION) USING THE SAME PRINCIPLES OF AERODYNAMICS.

A PROPELLER IS JUST A WING THAT ROTATES INSTEAD OF FLAPPING.

LARGE NUMBERS OF PHYSICAL SWITCHES (TRANSISTORS OR NEURONES) GIVE SUPPORT FOR ASTRONOMICALLY LARGE AMOUNTS OF VARIABILITY IN VIRTUAL MACHINES. N SWITCHES SUPPORT  $2^N$  POSSIBLE STATES, AND EVEN MORE STATE-TRANSITIONS OVER TIME. EVOLUTION "DISCOVERED" THIS TOO. (HOW?)

### VIRTUAL MACHINES IN THE BIOSPHERE

We can think of the biosphere as containing not only physical things like water, rocks and mountains, but also many concurrently interacting virtual machines, such as:

- different collections of genes
- different niches
- feedback loops between niches and gene pools,
- different information processing systems, in individual organisms, in groups and colonies, in ecosystems
- social groups
- poverty, crime, war, economic inflation, ...

Over time, designs develop and change, and the niches which influence designs also change, in a large web of interacting feedback loops driving the processes of evolution. These processes are all ultimately implemented in physical processes (e.g. molecular biology). They can also be perturbed by physical processes, e.g. climate changes, volcanic eruptions, continental drift, asteroids hitting the earth, etc.

Typically there are large numbers of niches and large numbers of designs all interacting in parallel, including cooperative and competitive co-evolution. Think of food chains. One organism's food link is another's waste disposal link.

Without powerful virtual machines in the biosphere could mere Darwinian selection among randomly generated "trial" organisms cope with the astronomical combinatorics in the time available? Hill climbing may not be enough to get from molecules to to man in a mere 4000,000,000 or so years.

**Compare: Stuart Kauffman** At home in the universe (1995)

### HIGHER LEVEL CONTROL THROUGH COGNITION

Development of organisms with cognitive abilities makes possible selection by detection of useful traits without an external designer.

- spartan elimination of "weaker" members
- choice of "stronger" mates

Are there other principles involved in interactions between co-evolving species (e.g. one species acting as a gradually more demanding "teacher" for another)?

There may be mechanisms which, once bootstrapped, accelerate certain types of evolution towards increased abilities, including increased information processing abilities. (Compare Kauffman on "autocatalytic networks" of molecules. )

Types of trajectories in design space and niche space: There are different sorts of trajectories possible for individuals, for groups of individuals, for designs and for niches. (Some trajectories involve discontinuous change.)

- I-trajectories POSSIBLE FOR AN INDIVIDUAL
- E.g. learning and development of individuals, or societies
- **E-trajectories** POSSIBLE ACROSS GENERATIONS, USING EVOLUTION **E.g. evolution of species**
- **R-trajectories** POSSIBLE FOR AN "EXTERNAL" REPAIRER

E.g. debugging of software, modification of a machine by a designer or repairer. These high level virtual machines are implemented ultimately in physics, via many different intermediate layers. They may exhibit principles of control, self-organisation, learning, etc. that we have not yet discovered.

Through bio-engineering new sorts of R-trajectories for organisms are being developed.

### CAN WE HOPE TO UNDERSTAND ALL THIS?

One way to impose structure is to explore the structure of "design space" the structure of "niche space" the relationships between them and the trajectories that can occur in those spaces, as individuals develop, species evolve, or machines are repaired or modified.

Instead of wasting time on fruitless debates about whether a particular type of mechanism (e.g. condition-action rules, theorem provers, or neural nets) is or is not necessary for intelligence we can explore the trade-offs in using or not using it, and analyse the effects of those trade-offs on options available during evolution.

In any case a neural net is just a collection of strongly interacting concurrent condition-action rules.

Slight differences in interactions between trade-offs and niches can cause divergence, so that different solutions develop in parallel. Geographical isolation can then cause further separation. Requirements for functional differentiation in social organisms may also promote divergence within a species.

#### An example trade-off – we find both:

systems where large numbers of small, expendable, individuals with relatively simple combinations of reactive behaviours suffice,
systems where there are relatively few, much larger, individuals with far more sophisticated information processing mechanisms, for instance planning and reasoning capabilities.

Compare ants and elephants. There are different social and individual information processing architectures.

# MAPPINGS BETWEEN DESIGN SPACE AND NICHE SPACE Between different designs and different sets of requirements



**Arrows depict different sorts of complex "fitness" relationships** (usually involving tradeoffs). Instead of fitness VALUES use fitness VECTORS.

Changes in one design can alter the niche of another, which in turn can lead to design changes, which alter the niche of the first. Interacting trajectories in both spaces may involve multiple feedback loops.

CONTRAST COUPLED FITNESS LANDSCAPES

#### **DIFFERENT KINDS OF TRAJECTORIES:**

I-trajectories POSSIBLE FOR AN INDIVIDUAL

**E-trajectories** POSSIBLE ACROSS GENERATIONS, USING EVOLUTION

**R-trajectories** POSSIBLE FOR AN "EXTERNAL" REPAIRER

# WHAT CAN BE INSIDE AN INTELLIGENT AGENT? A generic scheme



Rectangles represent short or long term databases, ovals represent processing units and arrows represent data flow.

Agents have various sensors and motors connected to a variety of internal processing modules and internal short term and long term databases, all performing various sub-tasks concurrently, with information flowing in all directions simultaneously. That still allows MANY variants, with MANY tradeoffs. In particular:

- architectures may be homogeneous or may have different types of mechanisms for different types of functions
- architectures may be "flat" or hierarchical or heterarchical
- there are many ways of achieving global decisions from multiple sub-processes.

### **REACTIVE AGENTS** HOW TO DESIGN AN INSECT?



#### IN A REACTIVE AGENT:

- Mechanisms and space are dedicated to specific tasks
- There is no construction of new plans or structural descriptions
- There is no explicit evaluation of alternative structures
- Conflicts may be handled by vector addition, simple rules or winner-takes-all nets.
- Parallelism and dedicated hardware give speed
- Many processes may be analog (continuous)
- Some learning is possible: e.g. tunable control loops, change of weights by reinforcement learning
- Agents cope using only genetically determined behaviours
- 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**



Some sort of "override" mechanism seems to be needed for certain contexts

#### AN ALARM MECHANISM:

- Allows rapid redirection of the whole system
- sudden dangers
- sudden opportunities
- FREEZING
- FIGHTING
- FEEDING (POUNCING)
- ATTENDING (VIGILANCE)
- FLEEING
- MATING
- More specific trained and innate automatic responses

#### Damasio and Picard call these "Primary Emotions"

## REACTIVE AND DELIBERATIVE LAYERS



#### 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 (The limbic system?):

Allows rapid redirection of the whole system

- Freezing in fear
- Fleeing
- Attacking (to eat, to scare off)
- Sudden alertness ("what was that?")
- General arousal (speeding up processing?)
- Rapid redirection of deliberative processes.
- Specialised learnt responses

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

### **SELF-MONITORING** (META-MANAGEMENT)

# DELIBERATIVE MECHANISMS WITH EVOLUTIONARILY DETERMINED STRATEGIES MAY BE TOO RIGID.

#### Internal monitoring mechanisms may help to overcome this if they

- 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, or using drawings to communicate about how things look.

## AUTONOMOUS 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
- Your self-evaluations may be ill-judged
- Your control may be partial (why?)

IT'S JUST ANOTHER TYPE OF PERCEPTION!

### "META-MANAGEMENT" PROCESSES MIGHT:

- Promote various kinds of learning and development
- Reduce frequency of failure in tasks
- Not allow one goal to interfere with other goals
- Prevent wasting time on problems that turn out not to be solvable
- Reject a slow and resource-consuming strategy if a faster or more elegant one is available
- Detect possibilities for structure sharing among actions.
- Allow more subtle cultural influences on behaviour

THE ALARM MECHANISM CAN BE EXTENDED

- Inputs from all parts of the system
- Outputs to all parts of the system
- Fast (stupid) reactions driven by pattern recognition

Imagine diagram extended: an octopus on one side with tentacles extending into all the other sub-mechanisms, getting information and sending out global control signals.

Humans seem able to learn to suppress some of these global signals. We can also learn to generate some of them voluntarily, e.g. in certain kinds of acting.

There's also a very complex chemical infrastructure with multiple subtle forms of long term and short term control (e.g. affecting mood, arousal, etc.).

#### **TERTIARY EMOTIONS (PERTURBANCES):** partial loss of control of attention. Possible only with meta-management which allows some control of attention.

Against Damasio & Picard: There could be emotions at a purely cognitive level – an alarm mechanism interrupting and diverting processing without going through the primary emotion system. We need to distinguish: (A) CENTRAL AND (B) PERIPHERAL SECONDARY EMOTIONS.

Both secondary and tertiary emotions may be either purely central or partly

peripheral.

### **INTERNAL COMPLEXITY**

### HIERARCHIC CONCURRENCY AND SPEED CONTROL

There are agents which act concurrently and asynchronously, but also components within individual agents which act concurrently and asynchronously, and components within components...

Relative speeds of different components may change (e.g. using resource-allocation strategies and architectures for dealing with problems due to resource limits, e.g. filters with interrupt thresholds, meta-management).

#### **COMBINING METHODOLOGIES**

Different \*types\* of mechanisms are likely to be required, including rule-based reactive systems, neural nets, parsers, meaning generators, sentence generators, pattern-directed associative knowledge stores, low level image analysers mainly crunching numbers, high level perceptual mechanisms mainly manipulating structures, simulations of other agents, event-driven and interrupt-driven modules etc.

Different sorts of agents with different sorts of architectures are geared to different tasks and requirements.

Many types evolved, chasing evolving niches. Individuals can also develop.

WHEN WE UNDERSTAND HOW ALL THESE DIFFERENT KINDS OF FUNCTIONALITY CAN BE COMBINED AND THE DIFFERENT WAYS THEY CAN BE IMPLEMENTED WE'LL BE BETTER ABLE TO ASSESS WHAT DIFFERENCE IT MAKES WHETHER THEY ARE IMPLEMENTED IN SOFTWARE, OR IN SPECIAL CIRCUITRY, AND WHETHER THERE ARE ANY SPECIAL ADVANTAGES IN THE PARTICULAR IMPLEMENTATIONS THAT EVOLVED NATURALLY.

### LEARNING AND SELF MODIFICATION

#### INCREASING ARCHITECTURAL COMPLEXITY INCREASES THE SCOPE FOR LEARNING AND DEVELOPMENT WITHIN AN INDIVIDUAL

• The more components there are, the more things there are that might be improved either by being self-adapting or via external mechanisms.

• The more components there are, the more scope there is for new links to be added, or for links to be modified (e.g. carrying richer messages).

• The more sophisticated the agent the more scope there is for improvements based on developing new representations. ARCHITECTURAL CHANGE

Individual agents, through learning or development, may need to be able to modify \*their own\* architectures, either to simulate biological processes of growth and development, or because applications of artificial agents require changes of competence at run time (e.g. agents extending themselves with new "plug-in" components at any level).

#### THINGS CAN GO WRONG IN MANY WAYS

- physical damage or chemical malfunction
- information distorted or missing
- wrong strategies or algorithms learnt or developed
- connections going wrong (missing links, or spurious crosstalk). Example: Multiple personality disorder.
- Some may be "fixed" by physical change.
- Some require "software change".
- Some may be irreversible.

### **THERE ARE MANY TRADEOFFS**

• Between transmitting information via genes (etc) and transmitting mechanisms for gaining the information

• Between having individuals "born" competent (precocial species) vs having them learn for themselves, possibly with parental help (altricial species)

• Between improvements through individual learning and development and improvements through social developments.

• Between having large numbers of simple (and expendable) individuals and having small numbers of larger and more sophisticated individuals

We don't yet understand all the tradeoffs and how they drove evolutionary processes, including the development of diverse solutions due to subtle interactions between tradeoffs and environmental differences (different niche pressures).

## CONCLUSION

We need a better understanding of how evolution explores and uses tradeoffs.

There are no "right" answers.

There are many discontinuities.

There are many hybrid solutions.

Many different breakthroughs, not yet understood, were needed for human architectures.

Yet insects remain pretty successful.

But that doesn't mean everything works like an insect ....

There are solutions in many kinds of "wetware"

Increasingly we are finding solutions in new kinds of hardware

Future solutions may exist entirely in software,

and be none the worse for that.