Slides for MetaCognition VideoPanel 19 Jun 2011
and EPICS Workshop 27 June 2011

Varieties of Self-Awareness and Their Uses in Natural and Artificial Systems

METACOGNITION AND NATURAL COGNITION

TOWARDS A CONCEPTUAL FRAMEWORK

(DRAFT: to be revised)

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

These slides will be added to my 'talks' directory:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/

Do Not Start from Definitions

I have seen huge amounts of effort wasted because researchers

– assume X is a good feature of systems

– try to define X

– try to build systems that satisfy their definition

DON’T START FROM DEFINITIONS:

DON’T ASSUME EXPERTS IN OTHER FIELDS KNOW MUCH ABOUT THE PROBLEMS

Most have no expertise in designing, analysing, debugging, explaining complex information processing systems.

START FROM REQUIREMENTS!

Arguing about definitions is generally a waste of time, and inventing your own definition, or even looking for a definition by some “authority” is a guarantee that you are likely to ignore something important.

There was a lot of that at the 2004 DARPA workshop on self-awareness.

`http://www.ihmc.us/users/phayes/DWSAS-statements.html`

How can you analyse requirements?

There's no simple answer – partly because there are so many different sorts of requirements

One way to identify requirements (of different types) is to look at products of biological evolution, attempting to understand the design discontinuities and

– the pressures that helped to select them

– including features of the environment

– their benefits, costs, flaws, limitations

It is often a fatal flaw to assume there is a single utility function

– i.e. some scalar measure to be optimised.

There are many incommensurable sets of requirements against which designs can be evaluated and often no right answer to "what is best"?

Compare consumer reports on products – cars, lawn-mowers, holidays, computers, ...

Herbert Simon and others: satisficing, not optimising is what natural systems do.

Probably that's all engineers can do, except for very simple classes of problems.

Sources of requirements: I

For artificial systems requirements analysis often involves

– examining some working or proposed system

– trying to identify flaws in existing systems

– interviewing people involved to find out what they like or dislike

– doing empirical research to find correlations between design features and results

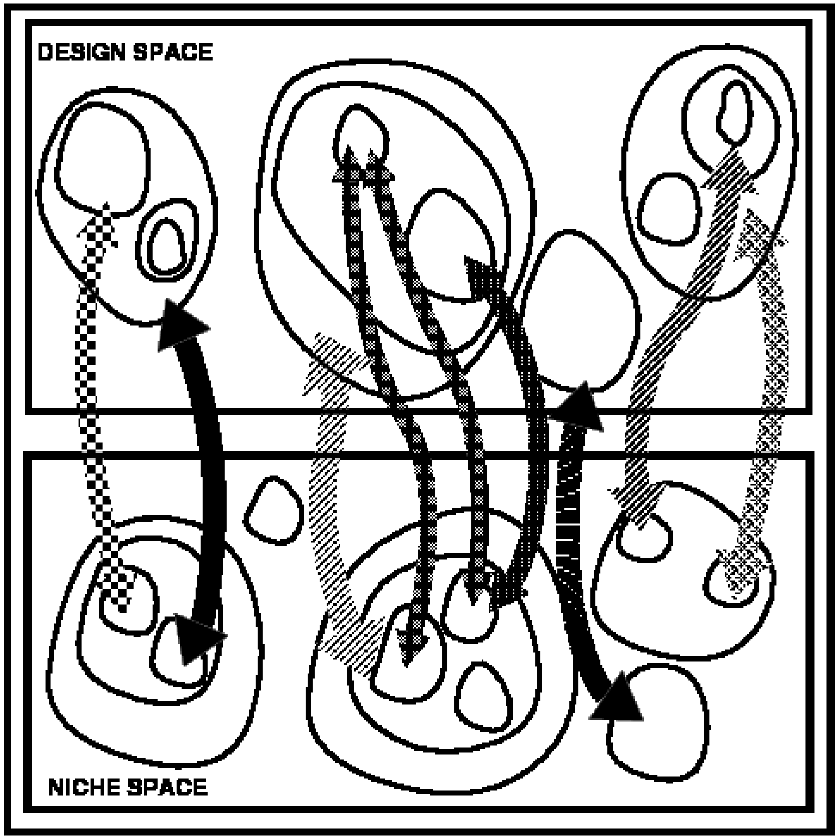
– often you don’t understand requirements until you’ve built working systems and found out what’s wrong with them

Sources of requirements: II

For natural systems we can try to understand trajectories in “niche space” and in “design space”

and how changes in both spaces interact.

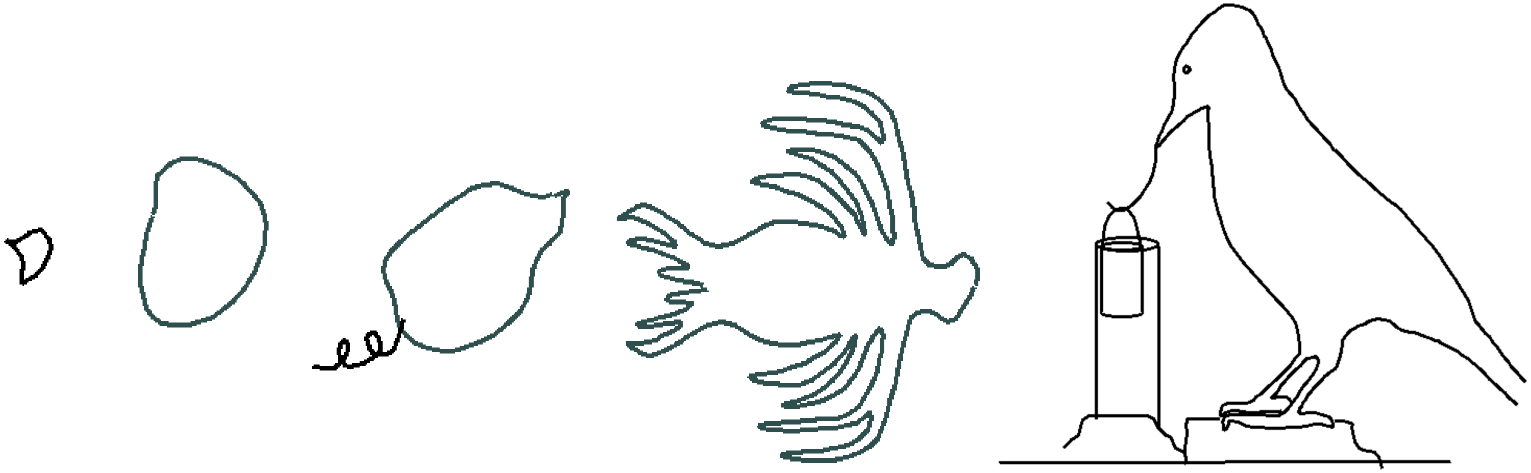
- Niche space: space of possible sets of requirements to be satisfied by working designs.
- Design space: space of possible sets of designs for working systems.



One approach is to try to identify design discontinuities in biological evolution and the factors that influenced them.

It’s not easy – partly because there are no fossil records of behaviour or information processing. Observing actual behaviour does not necessarily tell you what’s going on.

All organisms are information-processors but the information to be processed has changed and so have the means



From microbes to hook-making crows:
How many transitions in information-processing powers were required?

Contrast these transitions:

- transitions in physical shape, size and sensory motor subsystems
- transitions in information processing capabilities.

Fossil records don’t necessarily provide clues.

Environments have agent-relative structure

The environments in which animals evolve, develop, compete, and reproduce, vary widely in the information-processing requirements.

If we ignore that environmental richness and diversity, our theories will be shallow and of limited use.

In simple environments everything can be represented numerically, e.g. using numbers for location coordinates, orientations, velocity, size, distances, etc.

In more complex environment things to be represented include:

- Structures and structural relationships, e.g. what is inside, adjacent to, connected with, flush with, in line with, obstructing, supporting...
- Different sorts of processes, e.g. bending, twisting, flowing, pouring, scratching, rubbing, being compressed.
- Plans for future actions in which locations and arrangements and combinations of things are altered (e.g. while building a shelter).
- Intentions and actions of others.
- Past and future events and generalisations.

How can all those be represented?

But: simple environments are an unavoidable starting point for newcomers to the field, as long as they are treated as educational stepping stones to the real research.

Varied environments produce varied demands

Types of environment with different information-processing requirements

- Chemical soup
- Soup with detectable gradients
- Soup plus some stable structures (places with good stuff, bad stuff, obstacles, supports, shelters – requiring enduring location maps.)
- Things that have to be manipulated to be eaten (disassembled)
- Controllable manipulators
- Things that try to eat you
- Food that tries to escape
- Mates with preferences
- Competitors for food and mates
- Collaborators that need, or can supply, information.
- and so on .....

How do the information-processing requirements change across these cases?

Design space and Niche space
Reactive architectures

We can’t just think up ONE good design (any more than chemistry can make progress by studying just one molecule): biological evolution must have encountered many different problems that led to design modifications.

So we need to understand the space of requirements (niche space) and the space of possible designs (design space) and how those spaces relate, including what sorts of design transitions and niche/requirements transitions there are.

[Compare using the periodic table of the elements to impose some system on the basis of chemistry?]

That led (with help from Luc Beaudoin, PhD 1994 (Beaudoin, 1994) (Beaudoin & Sloman, 1993)) to a crude draft schema for architectures (CogAff).

The idea was built up in layers:

Earliest organisms had sensors and effectors and internal processing of various kinds of complexity, but could only react to the information currently available, without considering alternative possible states of affairs, past, present (e.g. remote), or future.

sensing $< -- >$ **internal processing** $< -- >$ **behaving**

That might suffice for micro-organisms in a chemical soup with perhaps detectable chemical gradients and items floating around that are either food (nutrients) or noxious.

Adding new architectural layers
Deliberative mechanisms

As the environment became more complex it became useful to store information beyond the instant of use (e.g. terrain maps and learnt generalisations).

This made it useful to add another layer of processing using different forms of representation including making use of more abstract ways of chunking information so that useful generalisations could be discovered and deployed.

These capabilities did not replace the old ones, which were still needed.

The new capabilties were added, providing two different “layers” of control: reactive and deliberative.

Deliberative: **sensing** $< -- >$ **internal processing** $< -- >$ **behaving**

many connections linking the layers
upwards, downwards, diagonally and sideways

Reactive: **sensing** $< -- >$ **internal processing** $< -- >$ **behaving**

For more on this see

Requirements for a Fully Deliberative Architecture (Or component of an architecture)

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604,

There are different kinds of self-awareness in different information-processing “layers”

Even examination of the value of a variable in a conditional instruction uses a kind of self-awareness.

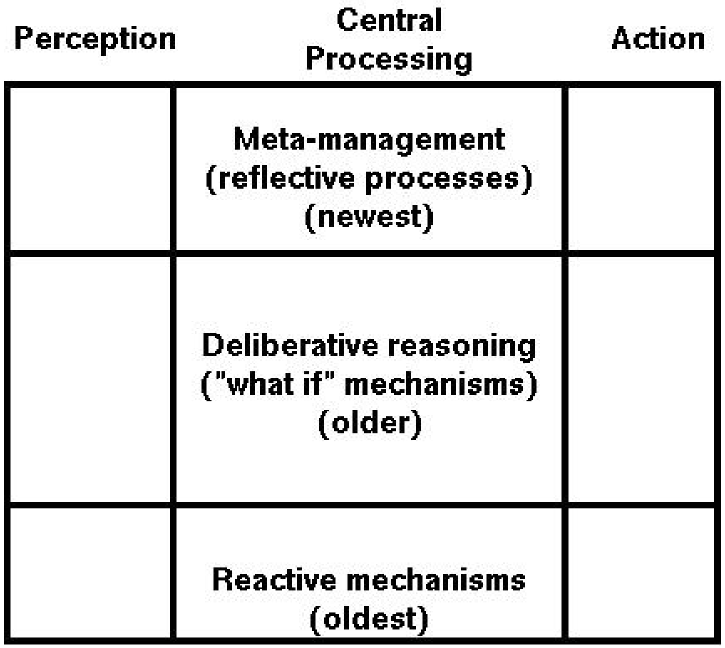
Many varieties of feedback and feedforward control

many other cases (Minsky, McCarthy, and others)

The CogAff Schema includes Meta-Management

The added complexity of information processing systems produced pressures for yet more abstract and sophisticated information processing capabilities referring not just to physical states and processes in the environment but also referring inwards to the forms of processing and types of information and types of goals and preferences that might or might not be useful or dangerous or ...

This added a third layer of processing called by some “reflective” and by us “meta-management”, to emphasise that it does much more than passive contemplation and analysis, but also actively manages things going on internally.



Perception	Central Processing	Action
	Meta-management (reflective processes) (newest)	
	Deliberative reasoning ("what if" mechanisms) (older)	
	Reactive mechanisms (oldest)	

That required new ontologies and new forms of representation (i.e. for representing things that represent). Those meta-semantic competences would also prove useful for perceiving, thinking about, interacting with, communicating with other organisms with or without similar capabilities.

The CogAff Schema covers a large variety of cases

With nine categories of components (boxes) generating a large variety of possible architectures depending on what sorts of mechanisms were in each box and which were connected with which others, etc.

[I think Minsky's six layers, in The Emotion Machine, can be viewed as sub-divisions in this framework.] (Minsky, 2006)

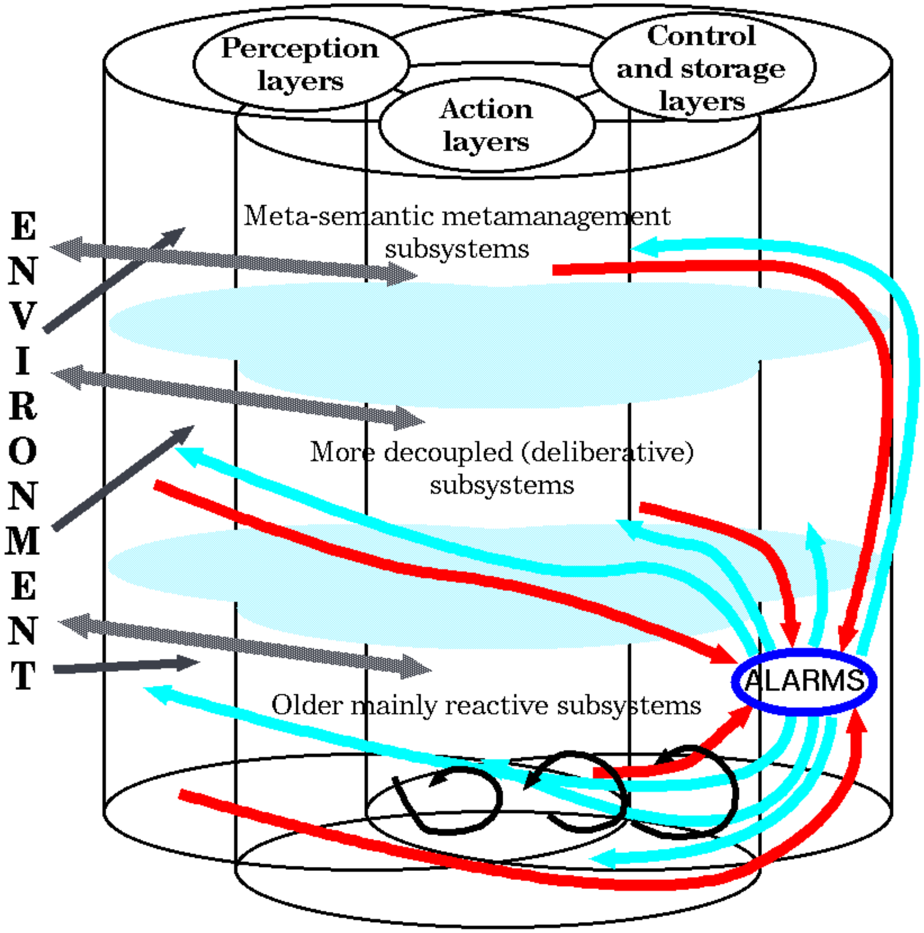
The CogAff schema leaves out lots of detail about types of motivation generation, and types of processing of motives of different sorts, many of which can be construed as metacognitive –

e.g. trying to determine which should be adopted, which rejected, which reconsidered later, etc. etc.

It also over-simplifies the subdivisions and overlaps between layers and boxes.

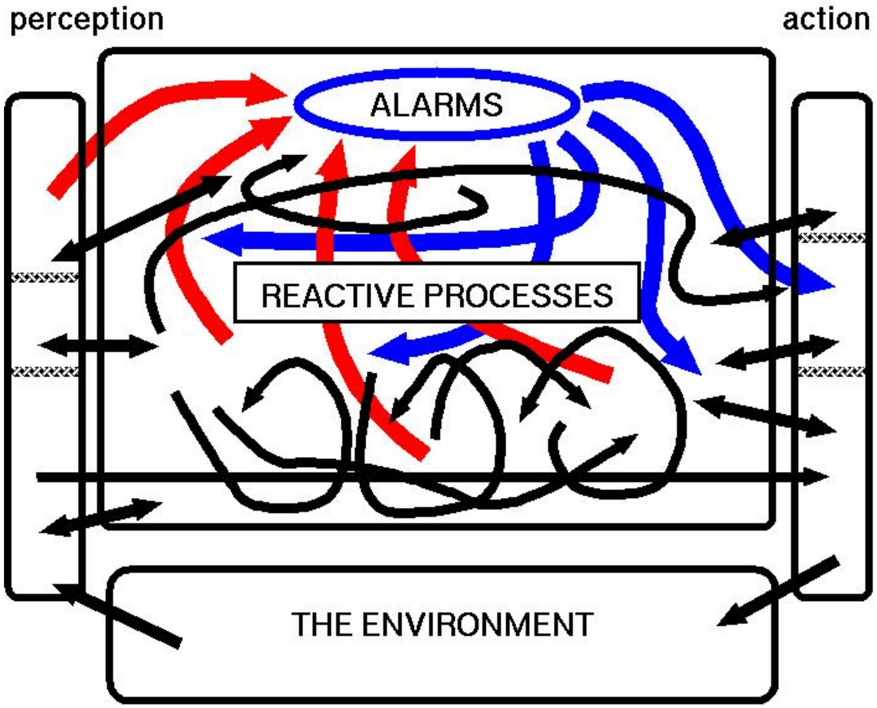
Another picture of the schema

Another (more messy) way of representing those boxes, emphasising overlaps between perception and action (as in J.J.Gibson 1966) is this (also showing a reactive alarm mechanism connected to non-reactive parts):


Perception layers
Control and storage layers
Action layers
ENVIRONMENT
Meta-semantic metamanagement subsystems
More decoupled (deliberative) subsystems
Older mainly reactive subsystems
ALARMS

There are many special cases of this schema, shown below.

Insect-like reactive architecture with alarm mechanism

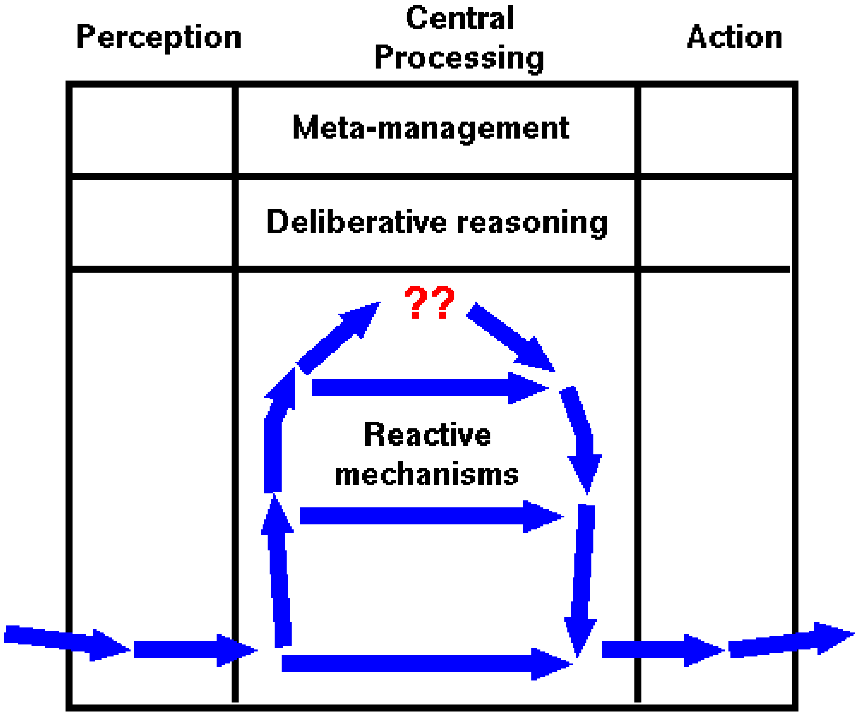


Insect-like reactive architecture with alarm mechanism

REACTIVE: not considering anything that does not exist in the immediate present, nor options apart from those immediately present.

Contrast: DELIBERATIVE mechanisms of varying complexity.

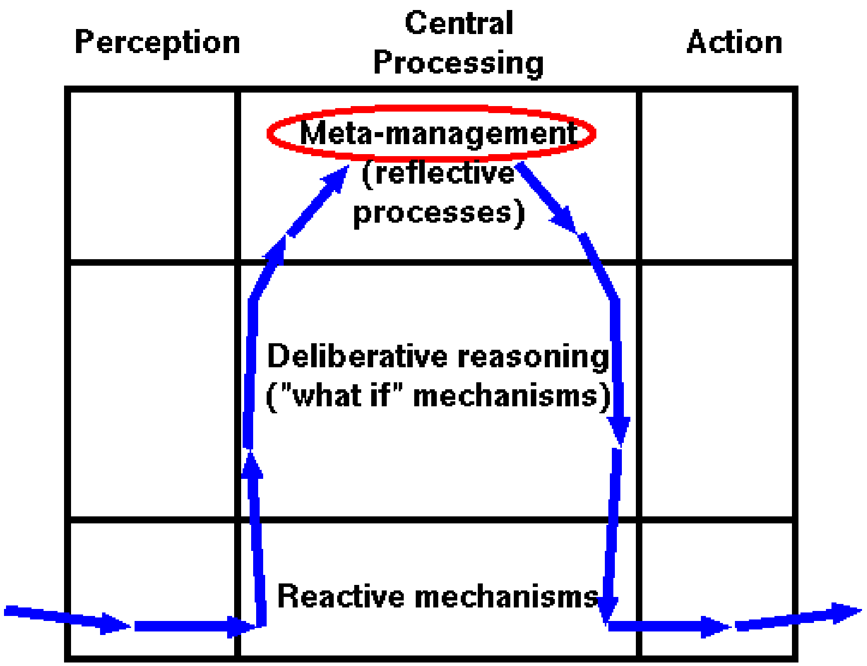
Brooks-like subsumption



Brooks-type subsumption (purely reactive)

Shallice and Cooper Contention Scheduling

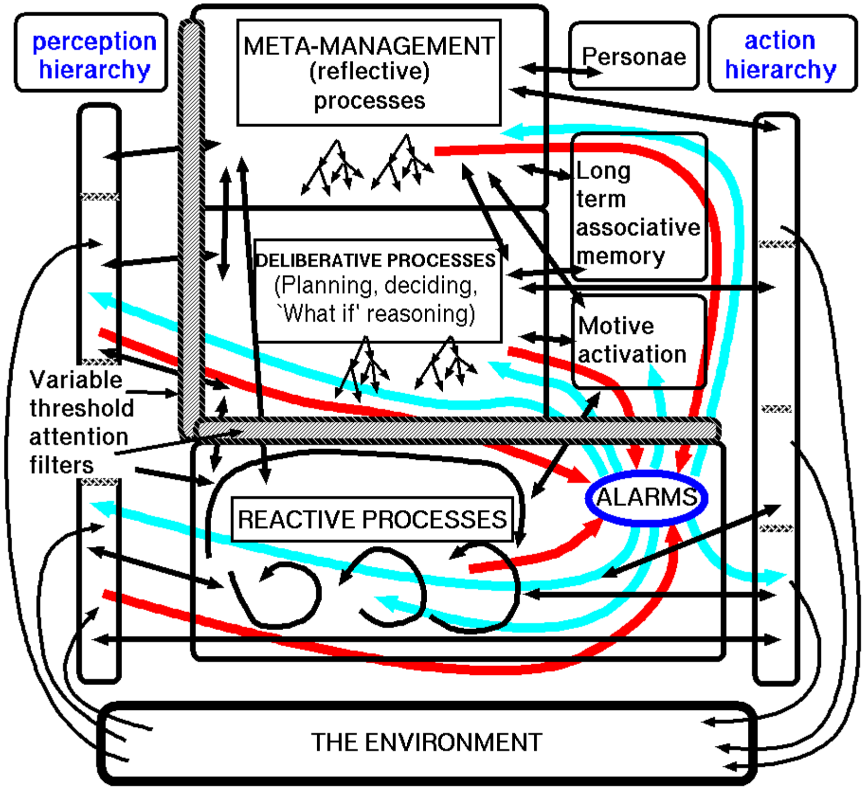
An “Omega” architecture – because of similarity to the Greek capital Omega.


Perception
Central Processing
Action
Meta-management (reflective processes)
Deliberative reasoning ("what if" mechanisms)
Reactive mechanisms

Shallice and Cooper (and Norman?) ”Contention Scheduling”
(more cross-links would make it subsumptive, but not reactive)

H-CogAff: Human-like CogAff architecture

NB This is unacceptably vague and conjectural: not definitive theory


perception hierarchy
META-MANAGEMENT (reflective) processes
Personae
action hierarchy
Long term associative memory
DELIBERATIVE PROCESSES (Planning, deciding, `What if' reasoning)
Motive activation
Variable threshold attention filters
REACTIVE PROCESSES
ALARMS
THE ENVIRONMENT

Sloman et. al. HCogaff (Human Cogaff: much simplified).
`http://www.cs.bham.ac.uk/research/projects/cogaff/`

There’s a VAST collection of possible designs with different amounts and kinds of metacognitive capability and different connections between the subsystems, requiring different physical mechanisms.

Many things need special treatment: e.g. language mechanisms are scattered around the architecture with their own special connections.

Architectures vs Architecture Schemas

CogAff is a schema H-Cogaff is a special case.

I suspect many of the architectures proposed by AI theorists, and also some proposed by neuroscientists, can fit into the CogAff framework, but they all use different diagrammatic and notational conventions, making comparisons difficult.

(Compare the BICA project: `http://bicasociety.org/cogarch/`)

Moreover, CogAff leaves open very many implementation details where others make choices.

I am mainly interested in trying to understand the subset of designs explored and used by biological evolution, as a way of understanding what the functions and constituents of human information processing might be, along with some other intelligent species (including elephants, corvids, primates, hunting mammals, octopuses, ...).

[

I don't know if all the possible designs can be implemented on digital computers.

Current physical machinery may not support all the required forms of causal interaction in virtual machinery. It's an open question, requiring much clarification.

]

Beware of confusions about embodiment and its importance

It’s not just a matter of what the organism is directly interacting with.

Compare:

what am I doing?

what else could I be doing, and what difference did it make?

what did I just do?

What didn’t I do?
Why didn’t I do it?
What else could I have done?

what am I going to do?

What other things could I do?
What would the consequences be?
e.g. what would I then need to consider doing?
(compare planning)

Warnings

NB: DO NOT ASSUME ALL RELEVANT STATE INFORMATION (OR CONSTRAINTS, etc) CAN BE EXPRESSED IN NUMBERS, VECTORS, ARRAYS, or EQUATIONS LINKING THEM.

What else is there?

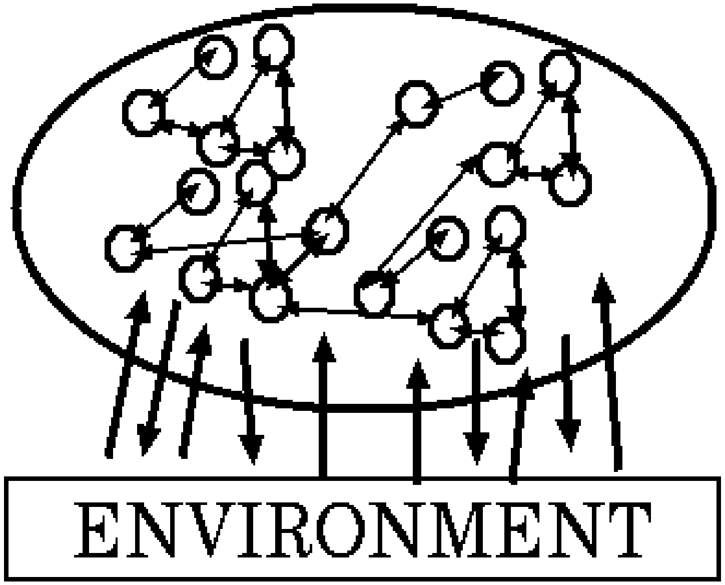
descriptions of structures, relationships, relationships between relationships, rules, grammars, trees, graphs,...

logical forms, collections of logical formulae, theories, deductions,

semantic relationships, epistemic relationships

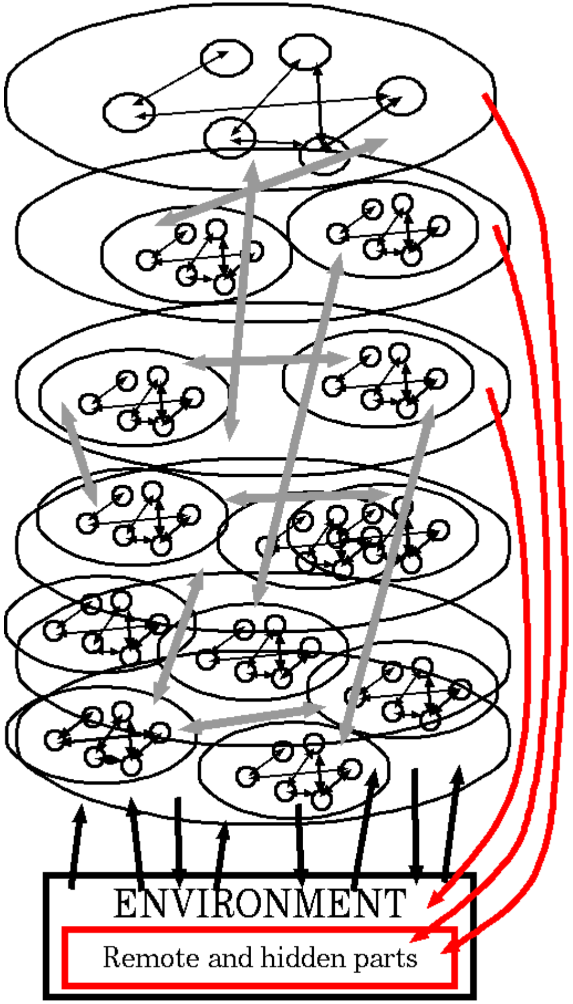
The dynamical systems view 1

Another way to view the same sets of possibilities involves collections of interacting dynamical systems, e.g. simple versions dealing only with the agent/environment interface (the short-sighted emphasis of most work on so-called ”embodied cognition”):



The dynamical systems view 2

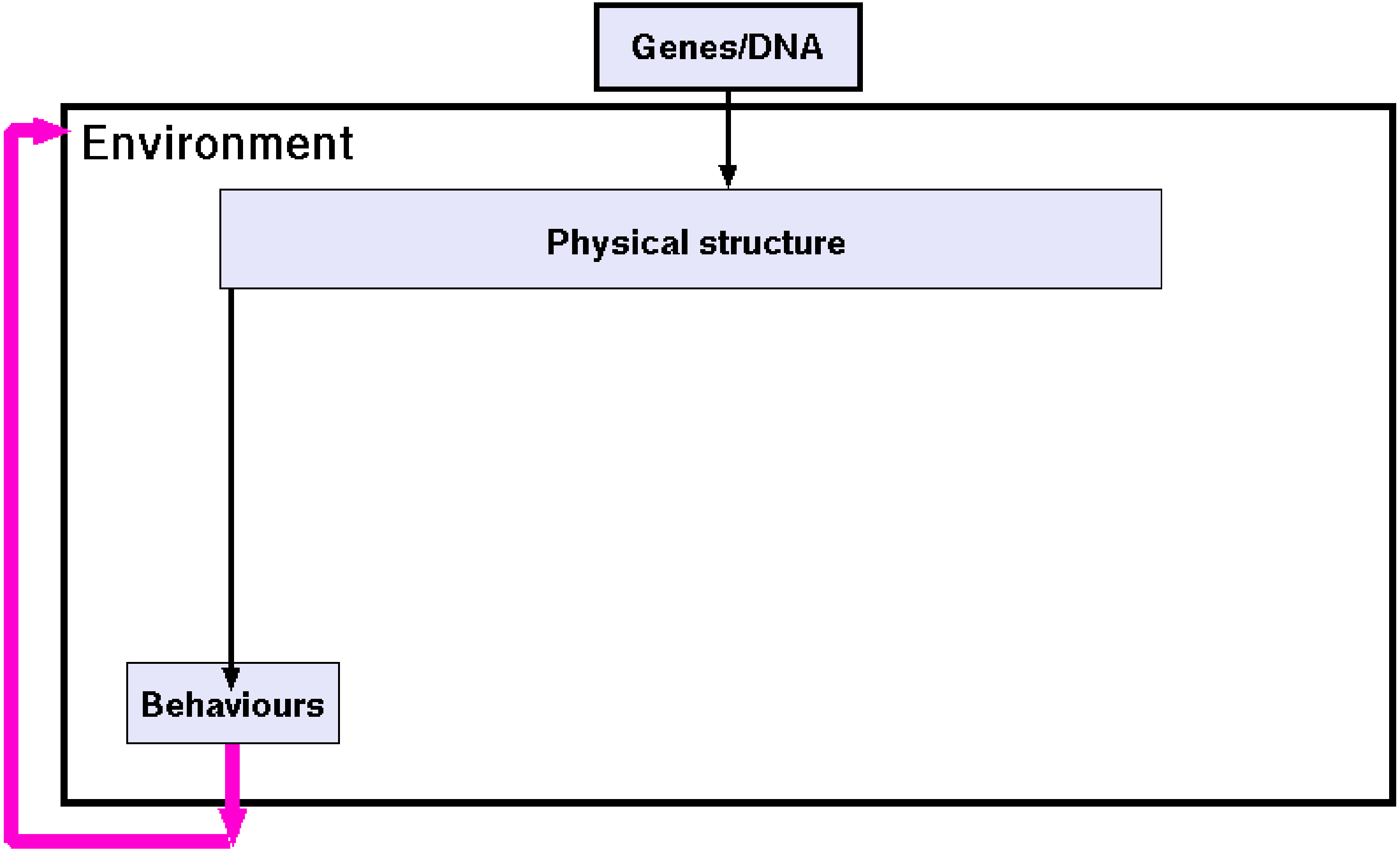
... and more complex multi-layered dynamical systems, grown by individual exploration and learning, some of them referring far beyond the immediate environment (e.g. studying history, or astronomy, or transfinite set theory – referring even beyond the physical universe).



We have barely scratched the surface of this exploration, but some interesting things have been learnt.

Epigenesis: Individual developmental trajectories

Routes from genome to behaviour : the direct model.

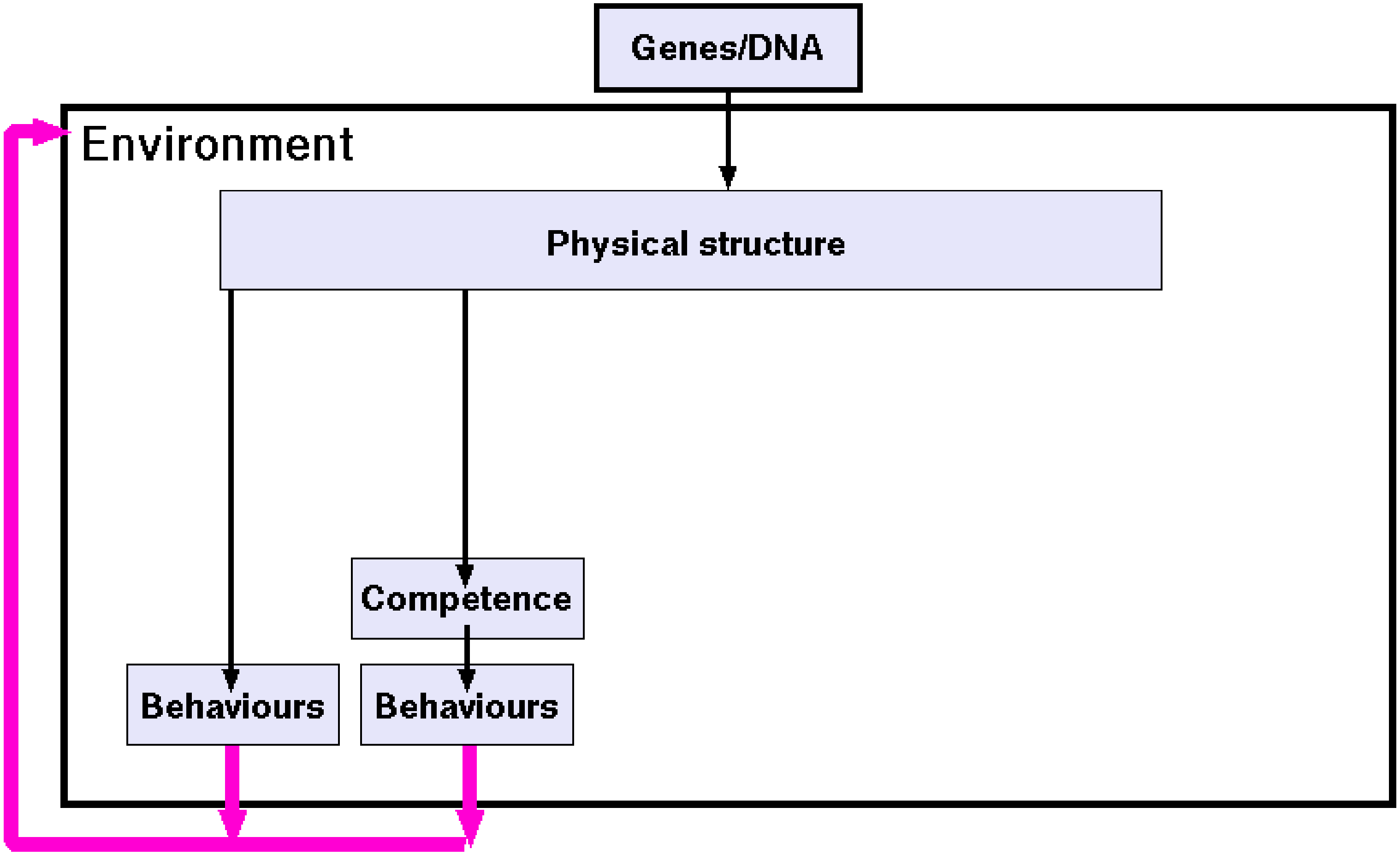


The vast majority of organisms (including micro-organisms) are like this.

Many don’t live long enough to learn much – they have to make do with innate reflexes. Other organisms have more “inside the box”.

Individual developmental trajectories

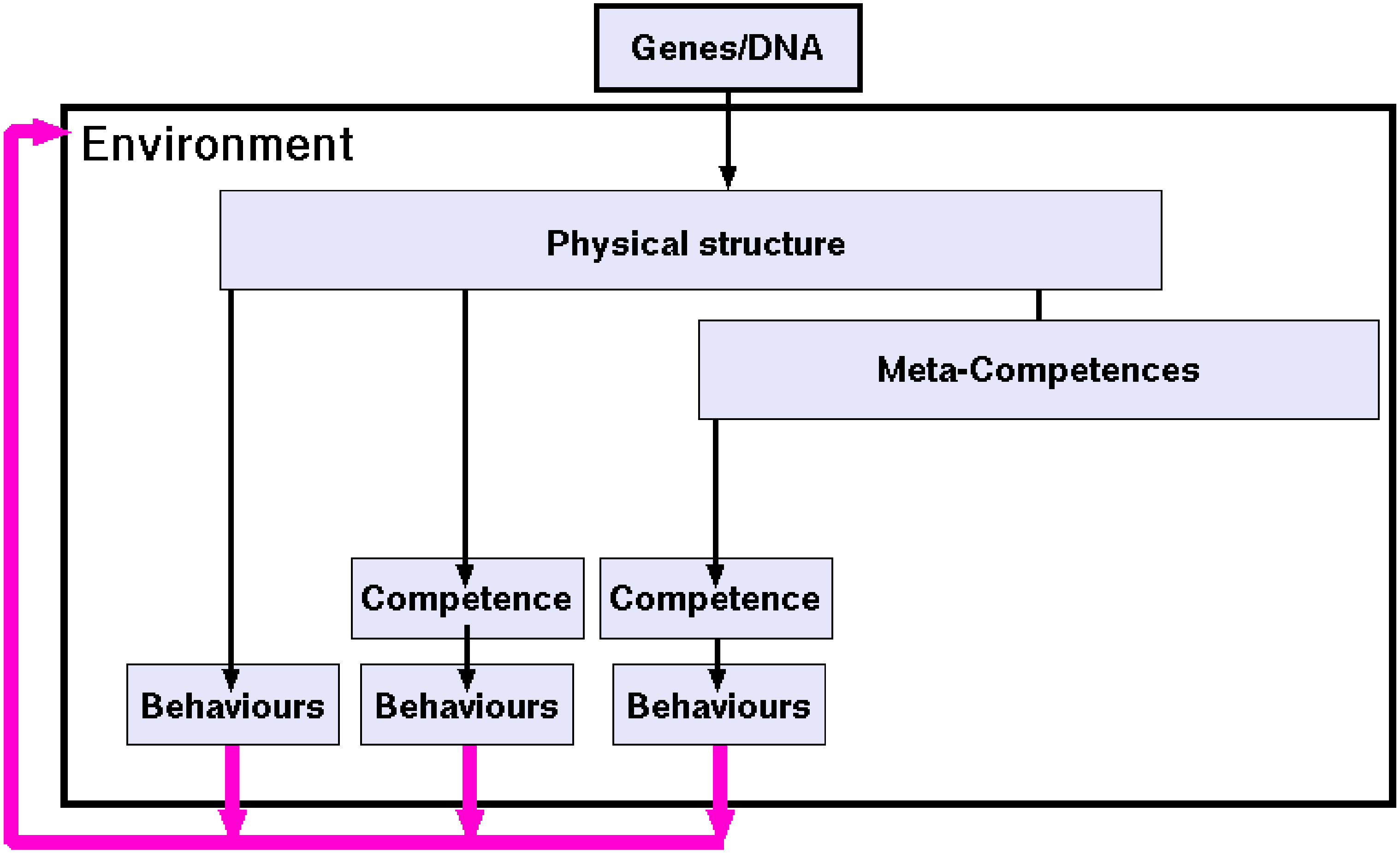
Routes from genome to behaviour : the two-stage model.



Some more complex organisms, instead of having only rigid (reflex) behaviours, also have competences that allow them to respond in fairly flexible ways to the environment: adapting behaviours to contexts.

Individual developmental trajectories

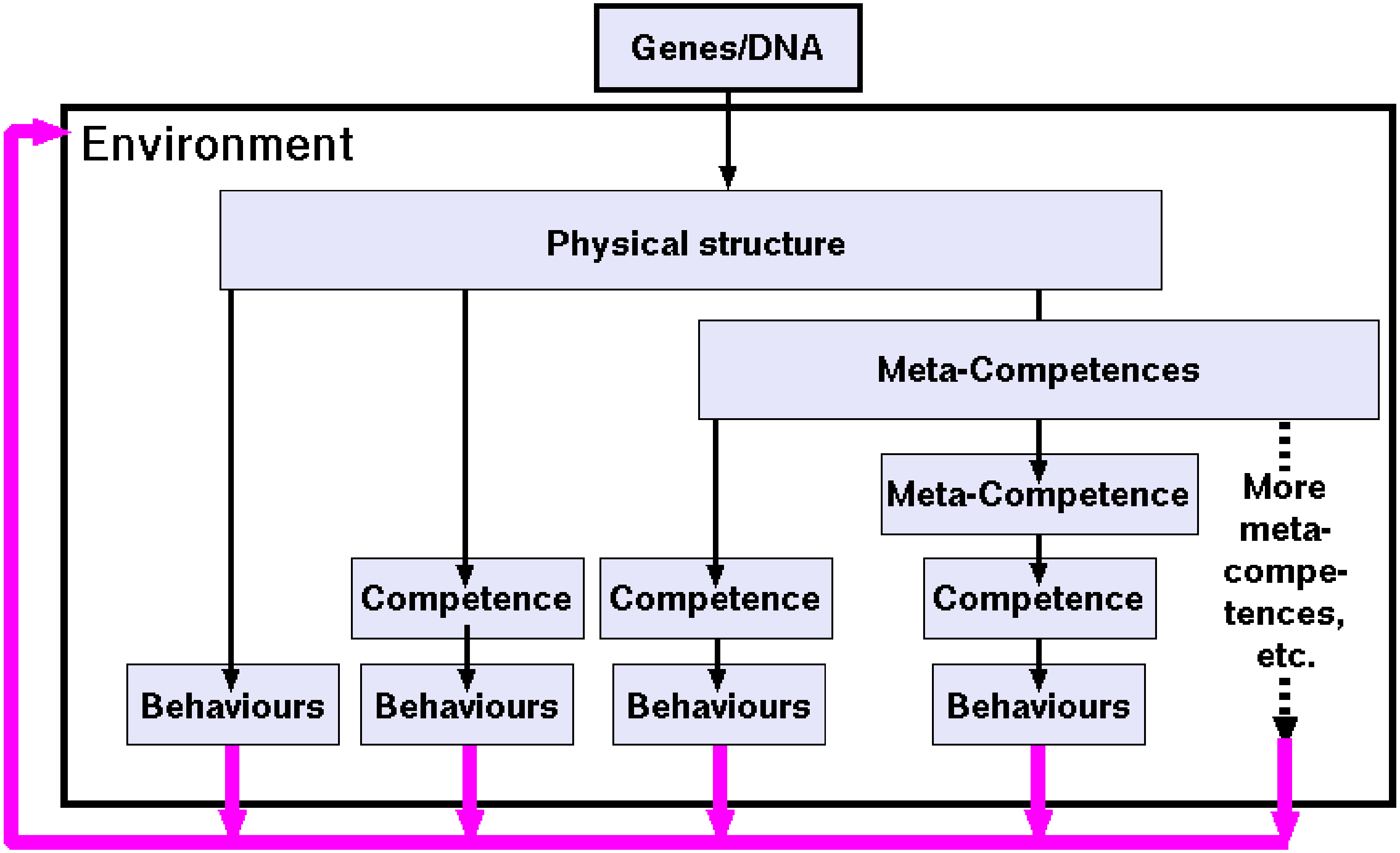
Routes from genome to behaviour : stages added by learning.



Genetically determined meta-competences allow individuals to respond to the environment by producing new types of competence, increasing flexibility and generality.

Individual developmental trajectories

Routes from genome to behaviour : the multi-stage model.



Some can also develop new meta-competences, on the basis of meta-meta competences.

Humans seem to be able to go on developing meta-meta-competences until late in life.

Implications of the Multi-stage Model for education

- **Children build their own information processing architectures**
  (Every now and again adding major new layers or subsystems.)
- Teachers merely help (but sometimes hinder) the process
  by providing building materials – along with challenges and opportunities.
  (Ideas from Jean Piaget, Ivan Illich, John Holt, Seymour Papert, Mitchel Resnick and others...)
- Learning by exploring and playing is a crucial aspect of human development.
  (And development in some other species.)
- Previous individual history limits what's learnable (e.g. when is a child ready for a new layer?)
- There cannot be a single learning trajectory followed by all children.
- We must find ways to let children (possibly in collaboration) build their own trajectories.
- **Vygotsky: Zone of proximal development:** what a child can learn at any stage is limited – but deep learning happens near the limits.
- Bruner and others: "Scaffolding" may be needed to support (and limit) the process.
- Only trivial things can be taught without generating confusion.
  Like building a map of terrain as you explore it – you will **inevitably** get things wrong at first and have to correct them later – as happens in the development of science and engineering.
  We need to understand that that's a feature of all deep education (e.g. learning a first language).
- Computers provide new opportunities for children and teachers to learn together.
  (Use of networks can extend the collaboration.)

Reminder: 1

**Don’t start
from definitions.**

**Instead:
Try to understand
sets of requirements
possible designs
and their relationships**

BigDog and LittleBoy are different.

Reminder: 2

If you want to do science:

Don’t present solutions

Without presenting
sets of requirements

And alternative
possible designs

and their trade-offs.

Merely showing something that works does not inform as much as showing things that nearly work, and explaining the differences.

References

References

Beaudoin, L. (1994). *Goal processing in autonomous agents*. Unpublished doctoral dissertation, School of Computer Science, The University of Birmingham, Birmingham, UK.

Beaudoin, L., & Sloman, A. (1993). A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, & A. Ramsay (Eds.), *Prospects for artificial intelligence* (pp. 229–238). Amsterdam: IOS Press.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Minsky, M. L. (2006). *The Emotion Machine*. New York: Pantheon.

to be extended