What Are Emotion Theories About?

AARON SLOMAN

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

School of Computer Science
The University of Birmingham

Related slide presentations can be found at
http://www.cs.bham.ac.uk/~axs/misc/talks/

Related papers: http://www.cs.bham.ac.uk/research/cogaff/

Tools: http://www.cs.bham.ac.uk/ axs/cogaff/simagent.html

Acknowledgements

This work was supported by a grant from the Leverhulme Trust
http://www.leverhulme.org.uk

Ideas presented here were developed in collaboration with

Steve Allen, Luc Beaudoin, Ron Chrisley, Stan Franklin, Catriona Kennedy, Brian Logan, Dean Petters, Riccardo Poli, Matthias Scheutz, Ian Wright

and others in the Cognition and Affect Project

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

Closely related work can be found at Marvin Minsky's web site, including his draft book The Emotion Machine, available at http://www.media.mit.edu/~minsky/

MORE THANKS

I am very grateful to the developers of Linux and other free, platform independent, software systems, including: LaTex, etc.

Diagrams are created using tgif, freely available from
http://bourbon.cs.umd.edu:8001/tgif/

Demos are built on Poplog

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

NOTE

The slides changed direction after discussions at the workshop. I started re-writing them to indicate the new directions, but don't know when I'll have time to finish, so here they are.

The revised bits are in slides 12–30 approximately

Precursors to validation

This was an invited presentation at the AAAI 2004 Spring Symposium workshop on

Architectures for Modeling Emotion: Cross-Disciplinary Foundations

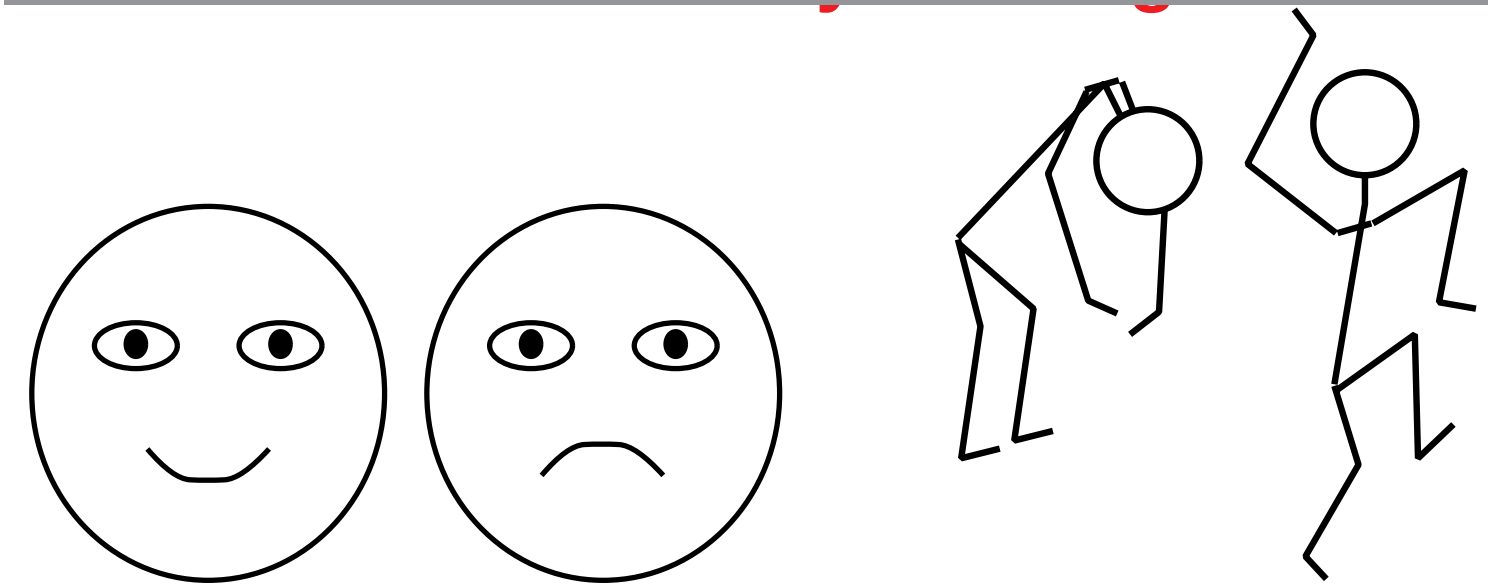
http://homepages.feis.herts.ac.uk/~comqlc/ame04

One of the concerns of the workshop was validation.

Validating a theory or model, or architectural description can either be concerned with practical usefulness, or with truth (or both). This presentation focuses on truth: truth as a theory of what something is and how it works, e.g. how human minds work.

THERE IS A PRIOR QUESTION

What do we mean by “having an emotion”?



- Is it enough to produce certain behaviours that people interpret as emotional?
- Do actors have the states they portray so effectively?
  e.g. jealousy, hatred, grief... not when such states include beliefs, intentions, as jealousy, hatred, grief etc., do.

- Behaviour is not enough to define any mental state, since
- In principle any behaviour, observed over any time period, can be produced by indefinitely many different mechanisms, using very different internal states and processes.
- We need to understand the variety of types of mental states better.
  Then we can define scientific concepts for classifying such states.

Asking whether a theory is true or false presupposes an answer to whether it makes sense at all.

- All theories use *concepts*,
- Insofar as the concepts are obscure, confused, or vague, the theories, and even the questions to which the theories are answers, will be flawed.
- Our concept of 'emotion' is highly ambiguous: an indeterminate collection of different possible concepts superimposed: even psychologists cannot agree on how to define it: there are many published inconsistent definitions of 'emotion'.
- This defect is inherited by questions and theories

  e.g. about how emotions evolved, what their functions are, which animals have them, which brain mechanisms produce them, what types there are, whether a foetus has them, whether they are all analysable in terms of "basic" emotions, etc. etc.
- Including the recent notion that intelligence requires emotions.

NOTE:

I am not saying that all relevant concepts have to be defined prior to theory-building: on the contrary the main way of making concepts more precise is to develop good theories using them. As theories about matter, energy, etc. developed in physics, the concepts became more precise, especially after Newton. Then Einstein changed them.

How can we produce theory-based clarification for "emotion"?

METHODOLOGICAL POINT

The concept of emotion is but one of a large family of intricately related, but somewhat confused, everyday concepts, including many affective concepts.

E.g. moods, attitudes, desires, dislikes, preferences, values, standards, ideals, intentions, etc., the more enduring of which (along with various skills and knowledge) can be thought of as making up the notion of a "personality".

Models that purport to account for emotion without accounting for others in the family are bound to be shallow

though they may have practical applications.

(See http://www.cs.bham.ac.uk/research/cogaff/talks/#talk3)

A "periodic table" for affective concepts can be based on an architecture, in something like the way the periodic table of elements was based on an architecture for physical matter.

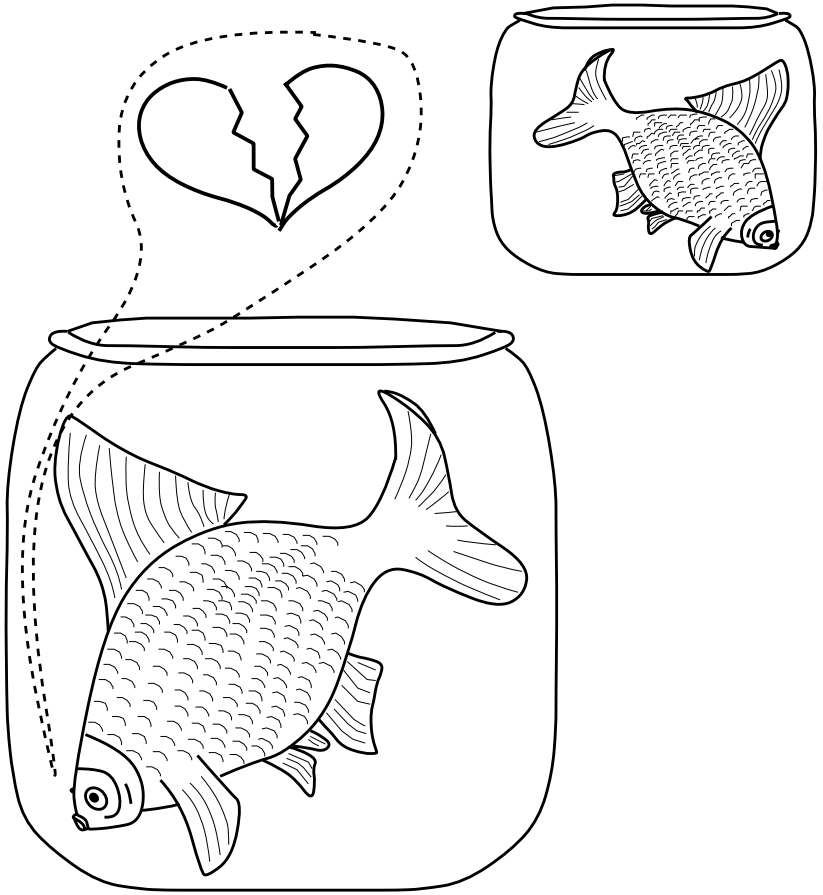
The analogy is not exact: there are many architectures for minds, each providing its own family of concepts.

So we need many periodic tables generating different sets of concepts.

There may be some concepts applicable across architectures

What's wrong with the concepts?

- Everyday concept of 'emotion' mixes up motivations, attitudes, preferences, evaluations, moods, and other affective states and processes.
- There's not even agreement on what sorts of things can have emotions
  - A fly?
  - A woodlouse?
  - A fish?
  - An unborn human foetus?
  - An operating system?
  - A nuclear power plant warning system?
- E.g. some people who argue that emotions are needed for intelligence are merely defending the truism that motivation is needed for action (though not in the case of tornadoes), and preferences are needed for selecting between options. Does a tornado select a direction to move in? Does a paramoecium?



Wishful thinking isn't science

Sometimes over-generalising the notion of 'emotion' is related to a *desire* to argue that emotions are important in ways not previously acknowledged

- If 'emotion' is construed so broadly that it covers all goals and preferences the claim that emotions are needed for intelligence is vacuous.

  Compare David Hume's claim that reason has to be the slave of the passions: without motivation, no amount of knowledge or reasoning power can produce action.

- If 'emotion' refers more narrowly to the sorts of processes in which one sub-system interferes with, disrupts, modulates or aborts the 'normal' functioning of another, as happens in many of the states in which people are described as being 'emotional' (e.g. infatuation, terror, obsessive jealousy, being startled by a movement) then it is false that emotions are *required* for intelligence:

  Emotions of that sort can get in the way of sensible decisions and actions.

- Saying that states of type X can occur as a side-effect of the operation of some mechanism M that is required for intelligence does not imply that states of type X are themselves required for intelligence.
  All it implies is the unsurprising claim there are some common sub-mechanisms shared between different functions.

Fallacious inferences

- **Damasio's widely quoted reasoning (1994) from the premiss:**
  Damage to frontal lobes impairs both intelligence and emotional capabilities
  to the conclusion
  Emotions are required for intelligence
  is fallacious.

- **A moment's thought should show that two capabilities could presuppose some common mechanisms without either capability being required for the other.**

- **A research community with too much wishful thinking (because people have not learnt to think as designers?) does not advance science.**

I, for one, don't need what I think most people call "emotions", most of the time

It has become highly fashionable in some circles to claim (and sometimes to argue – though invalidly) that intelligence needs emotions.

Written around 3.a.m. Monday 22nd March 2004:
I hope I'll finish preparing these slides in time for the meeting.

But I am not at all emotional about this.

I am not even really anxious.

If I don't finish I won't, and I'll have to do some extemporising.... not for the first time..... and sometimes that produces a better talk.

I hope I am doing this intelligently, despite not being at all emotional about it. I am, of course, motivated, but that does not imply being emotional.

Tornadoes and tadpoles

What's common to and what's the difference between a tornado and a tadpole, between a monsoon and a mouse? ...

Tornadoes and tadpoles

What's common to and what's the difference between a tornado and a tadpole, between a monsoon and a mouse? ...

ANSWER:

How energy is deployed:

Some things use information to deploy energy to meet their needs (etc...) and others just act.

Let's start again: using the commonalities and differences between biological organisms from the simplest to the most complex, in an attempt to understand the design principles "discovered" by biological evolution – instead of trying to understand the hugely complex human case in isolation from the rest.

DON'T TRY TO RUN BEFORE YOU CAN WALK.

DON'T ENTHUSE ABOUT SOME PARTICULAR AI ARCHITECTURE.

WE NEED TO UNDERSTAND GENERAL PRINCIPLES.

Towards a general theory of affect

There are some high level features common to all living things, but absent in rocks, clouds, rivers, stars, atoms, marbles, and so far missing from any artefacts we have made (though there are some *partial* replications)

Let's assume the following for now, and see where it leads us:

- **Organisms have needs (things required for survival, reproduction, repair, ...)**

  We can also talk about inorganic things having needs, e.g. a painting, a car, even a rock:

  "That rock needs shoring up, to stop it falling apart after the damage caused by weathering."

  Any necessary condition for something can be described as a need!

  But inorganic things (except for some artefacts) don't satisfy the following descriptions:

- **Organisms acquire and use information**
  - **to select actions that serve their needs,**
  - **to control the details of the actions**
  - **using energy available to the organism in performing the action,**
    - **usually chemical energy stored in the organism,**
    - **but sometimes other sources of energy available in the environment, e.g. wind-power, water power, gravitational potential energy.**

- **Some of the information is *control information*, and some some *factual***
- **Organisms have sub-mechanisms performing different sub-functions, using different information expressed or encoded in different forms.**

KEY IDEA:

By developing a systematic theory of systems that acquire and use information for control we can give our concepts for describing states and processes in information-processing systems improved clarity, precision, coverage of cases, usefulness in explanatory theories, ...

We use "affective" as a very general label for control states whose function in an architecture is to initiate processes, prevent things from happening, abort, redirect, suspend, speed up, slow down, prioritise processes etc.

Without some affective states in an organism or machine nothing will happen, no matter how much factual information it has, and no matter how many planning, reasoning, problem-solving or perceptual capabilities it has. In addition to this most basic and general notion of 'affect' we can introduce many special cases.

'Affective' is a technical term. Our use is very general. Many people restrict it to states in humans that include hedonic feelings, e.g. pleasure and pain – but those states require very rich and complex architectures, whereas for a general theory of architectural design for organisms we need some general terms applicable even to the simplest organisms.

Later we'll characterise many sub-types of affect, including types that correspond to some of the ordinary, non-technical, kinds of uses of the label "emotional".

NOTE ON TERMINOLOGY

We are using many terms for which scientists and others already have some intuitive understanding but which require analysis that will not be provided here, e.g. 'need', 'function', 'mechanism', 'architecture', 'information', 'representation'.

For more details see this presentation

http://www.cs.bham.ac.uk/research/cogaff/talks/#inf

and papers in the Cognition and Affect project directory,
e.g.

A. Sloman, (1993), The mind as a control system, in *Philosophy and the Cognitive Sciences*, Eds. C. Hookway & D. Peterson, Cambridge University Press, pp. 69–110

A. Sloman, R.L. Chrisley & M. Scheutz, (To Appear), The Architectural Basis of Affective States and Processes, in *Who Needs Emotions?: The Brain Meets the Machine*, Eds. M. Arbib & J-M. Fellous, Oxford University Press, http://www.cs.bham.ac.uk/research/cogaff/sloman-chrisley-scheutz-emotions.pdf,

Towards an ontology of architectural 'features' and 'feature-extensions'

- Over billions of years, evolution found more and more diverse, increasingly sophisticated, "design principles" relevant to making things better able to meet "top level" needs to survive and reproduce.
- As complexity increased and the environment (including other organisms) produced more and more challenges, new niches produced new design problems requiring new design principles for their solution, leading to ever more diverse and complex architectures and mechanisms.
- We cannot hope to retrace the whole of evolution, but perhaps we can work out with hindsight what many of the design problems were, and produce a taxonomy of useful design principles for addressing those problems.
- Since the 1950s we have learnt very many design problems and principles relevant to information-processing systems, expressed in increasingly sophisticated programming languages and formalisms for specifying software systems.
- But most of that work was not concerned with requirements for self-motivated, self-maintaing, self-repairing, reproducing organisms, with concurrently active sub-systems.
- We can still use much of what has been learnt in AI. Software Engineering and Control Engineering even if we re-start with a biological viewpoint.

Lessons from “simple creatures”

By considering hypothetical creatures of various sorts (starting with simple cases) we can investigate general design issues

(Compare: V. Braitenberg, (1984), Vehicles: Experiments in Synthetic Psychology, MIT Press)

- Every organism needs motivators – e.g. devices that detect some need (e.g. a need for nourishment, or temperature adjustment) and attempt to initiate appropriate action, e.g. by sending control signals to motors, or to other internal mechanisms. Motivators vary in several ways:
  - Sensed states and motor signals may vary continuously or discretely
  - Control may be ballistic or online, using feedback (e.g. homeostasis)
  - Action required may be determined by need or may require additional sensors whose results are used to select between possible actions according to context (e.g. in winter dig for roots, in summer climb trees for fruit)
- Most organisms have two or more (usually more) concurrently active motivators. This can generate conflicts: two needs require different actions. Many resolution strategies are possible, e.g.
  - first motivator activated wins
  - activations are combined (e.g. vector sums – usually a bad idea)
  - a ‘comparator’ mechanism selects one motivator: winner takes all
  - needs are met sequentially (first get food then get drink, etc.)
  - a creative compromise meets more than one need: ignore nearby food and drink sources (in opposite directions) and go in a third direction to a more remote location providing both
- Motivators may need to share resources,
  e.g. sensors (for checking conditions) and motors (for moving to achieve goals).

Example: simple affective (desire-like) states.

Consider an organism O with two sensors SF, SP.

- **SF measures an internal state (e.g. an energy-level) and if that is below a certain threshold value, O initiates food-seeking behaviour.**
- **SP detects proximity of predators and if one comes nearer than some threshold, O generates escape behaviour.**

In this case we can say that the states of sensors SF and SP express control information: their states are desire-like, or motivational states of O).

In principle motivators concerned with different needs can generate conflicts, e.g. if SF and SP generate inconsistent behaviour.

In some designs no conflict would arise.

E.g. assume that if one or other behaviour has already been initiated, the sensor for the other behaviour has no effect, though if the initiated behaviour terminates, the other sensor becomes capable of initiating behaviour.

This would probably not be a good design!

An alternative would be always to allow one motivator to dominate the other, e.g. FP always wins.

There are many other 'conflict-resolution' mechanisms.

Proto-deliberative states

Consider an addition to the organism O: something that receives and compares inputs from the two sensors SF and SP, namely information about the amount by which each threshold has been passed, or zero if the threshold has not been passed.

- **If neither of SF or SP has passed its threshold the comparator does nothing.**
- **If only one of SF or SP has exceeded its threshold, the comparator simply passes on the signal from the active sensor to the motor mechanisms to produce food-seeking or predator-avoiding behaviour, as before.**
- **If both have exceeded their thresholds, the comparator passes on the signal from the one with the greater excess, suppressing the other — i.e. it uses a 'winner-takes-all' mechanism. (This assumes the scales are commensurable!)**

We can describe this as a 'proto-deliberative' mechanism. It implicitly 'considers' two options and selects one.

It is merely proto-deliberative, because, unlike fully deliberative systems, it cannot explore and compare multi-step futures, as planning mechanisms typically do.

If the comparison process takes some time, we can say that while the comparison has not yet been completed, the organism O is in a 'state of indecision' of a rather primitive sort. (This does not require O to be aware of being undecided.)

Example: mechanism for tied-conflicts

- Suppose that in our last example the time it takes for the comparator to reach a decision depends on the difference between the two measures, so that if they are very close 'settling down' takes a long time.
- In that case it might be useful to have another device, CZ (the Comparator Zapper) with a clock which gets turned on whenever the comparator starts attempting to reach a decision.
- If the time taken exceeds some threshold, then CZ overrides the comparator and selects one of the options at random.
- Most people will recognize this design problem and the corresponding solution as relevant to much more complex, human-like systems!
- This is one of many contexts in which it is useful for one process to interrupt, modulate, suspend, restart, or otherwise interfere with another. We could call such processes interventions. Many of the things we call emotions seem to have the character of interventions, though they vary in many details.

  E.g. being startled is a very short term low-level intervention, whereas being infatuated with someone is a relatively long term intervention that can modify or distort many processes of different sorts.

  Compare: H. A. Simon, (1967), Motivational and emotional controls of cognition, Reprinted in *Models of Thought*, Yale University Press, 29–38, 1979

Factual, belief-like, information states.

- Consider an organism like O with the sensor SF as before. Suppose there are two sources of food: one where there is high quality food but daylight is needed to find it, and another were the food is poor, but easily found in the dark.
- Then it will be useful to have another sensor, SI which reacts to illumination, and whose output does nothing on its own, even when it changes, but which is used by SF when it detects a need for food (above its do-nothing threshold).
- Then instead of directly triggering motion, the food-seeking motivator might trigger motion in one direction or another depending on the information provided by SI.
- We can then say that the state of SI is a factual or belief-like state, used by mechanisms processing desire-like states.
- The same factual information could be used by several different motivators.

  E.g. the action triggered by SP (predator detector) might be to flee in good light and to freeze in bad light.

- If output of SI varies continuously with illumination, then SF and SP will each have to apply a threshold in order to select an action.

  If the same threshold is useful for both SF and SP it may be be more economical to have a single threshold mechanism in SI so that its output is essential a boolean: illumination is either high or not high.

A meta-level motivator

- If SI is a thresholded-illumination detector whose output is used by several different motivators, it could be important for SI to use different thresholds when used by different motivators:

  E.g. the level of illumination at which it is safer to flee from predators instead of freezing may be different from the level which separates the two feeding strategies.
- In that case it may be useful to have a device which senses whether SF or SP is currently active and sets the threshold in SI accordingly.
- This new device can be thought of as a motivator whose purpose is to set SI's threshold so as to serve the needs of SF and SP.
- Its sensing state expresses a desire-like state which triggers only internal actions.
- An alternative solution would be to make SI a parametrised thresholder: any device that interrogates it provides a threshold level as a parameter, which determins the answer it gets.
- This is another area where conflicts could arise between the needs of different sub-systems using the same illumination detector. As always, there are very many possible designs for arbitration mechanisms.

A factual sensor can become an affective device

Suppose O's illumination-sensor SI produces a result that causes food seeking behaviour suited to high illumination: what should happen if after the behaviour has begun the illumination crosses the threshold beyond which the other food seeking option would have been chosen. What should happen? There are various options:

- **There is no effect, and food-seeking continues as before.**
- **There is only an effect if the change is from conditions for going to poor food to conditions for going to good food.**
- **There is an effect, but only if the change happens soon after the original decision was taken (so that O has not got far in the no longer optimal direction).**

In the cases where the change in SI triggers re-direction of food-seeking, this is another example of an intervention, and the sensor that previously provided only factual information now has control powers and its state therefore fits the requirement for being affective.

So the very same mechanism can be sometimes non-affective and sometimes affective (even "emotional") in its functioning.

Note that very similar behavioural effects could be produced by increasing the complexity of the food-seeking behaviour, by building into it frequent checks on illumination with appropriate rules, one of m any cases where "emotionality" (i.e. intervention) can be useful, but is not the only option.

Implicit and explicit information states

So far we have discussed only implicit, not explicit, information

- **Implicit information states:** factual or control information states in which the information exists only while it is being processed (i.e. being caused by its typical cause and causing its typical effects, e.g. states of SI, SF, SP)
  - The state of a sensor or array of sensors while sensing something
  - The state of a goal or action-producer while it is actually producing (or, in the case of conflict, tending to produce) the action.
- **Explicit information states:** states bearing factual or control information where some structure or sub-state persists as a *record* or *summary* of the information.
  - Explicit information can endure beyond its typical cause and beyond its typical effect
  - It can even anticipate the implicit states ('I am going to be hungry' '...to hear a dinner bell')
  - It can be operated on by mechanisms that manipulate symbolic structures, e.g. combining them (in predictions, generalisations, plans,...)
  - It can be *communicated* between sub-systems, or between complete individuals.

  Some people promoting "dynamical systems" approaches want us only to consider implicit information states – swirling bathwater, never mind the baby.
- The use of explicit information states – enduring, manipulable, "symbols" can serve many biological needs in a powerful way. HOW?
- Our simple organism O has no explicit information states. It cannot record a state it was in previously or might be in.

  ("Implicit" does not imply unconscious: we can be aware of implicit states and unconscious of some explicit symbols, e.g. in linguistic processing.)

Evolutionary vs developmental or learning trajectories

We have presented several examples of a change from one design for an organism (or sub-system) to another design that provides new benefits (and perhaps costs).

How do such changes arise? There are several possibilities.

- **Evolution:** The change could occur across generations, via natural selection.
- **Development:** The change couuld occur because the genome already has the option for various changes, which are triggered by aspects of the environment encountered by O e.g. patterns of success and failure.
- **Learning:** The change could occur because O has a general learning mechanism which searches for ways of changing behaviours of sub-systems that produce new benefits, such as energy-saving, fewer injuries, etc.
- **Repair/re-design:** In the case of an artefact the change could occur either within an individual, or across 'generations' or 'versions' of the same product because an engineer alters the design. If the change is made to a running system, O, this may or may not require O to become temporarily non-functional. (Cf. 'plug-ins'.)
- **Social transmission or teaching:** In some cases the change does not require physical modification to the individual because it already has a design with so much flexbility that it can absorb information from others about how to change itself: this could involve explicit instruction, or implicit cultural transmission, sometimes describable as indoctrination.

Some issues re-appear at different levels of complexity

Norbert Wiener pointed out in his 1948 book that there are several general principles of control that are applicable to systems which differ in in physical structure, in complexity, and in what they do — one of the most familiar being the use of negative feedback to achieve homeostasis.

- This general applicability is also true of the design issues we have discussed: e.g.when talking about a light sensor we did not specify its physical design, whether it was waterproof or not, how big or small or accurate it was, etc.
- Likewise predator detectors and detectors of need for food can vary enormously in their physical design their complexity, their size, etc.
- The effectors/motors that are activated by them can also vary enormously, and yet the design principles and options already mentioned could apply to any of them.
- Likewise the need for intervention can occur at many levels and in many forms, with very different detailed requirements.
- Some of the distinctions can become fuzzy in places: e.g. a feedback control loop has continuous 'intervention' though it may not be useful to think of it as anything like an emotion because such control mechanisms are so commonplace and because the intervention is continuous and part of a mechanism's normal operation, rather than an occasional disruption.

Increasing complexity: modulations of and interactions between control states

We have illustrated some simple ways in which new functionality can be added to an existing architecture, increasing the "periodic table" of possible states and processes associated with the architecture. They illustrate a number of general principles – though many more remain to be studied.

- **Some additions allow a continuosly varying information state to be discretized, to support new functionality.**
- **Some additions provide extra information for a pre-existing control mechanism, making it able to tailor actions more precisely to circumstances.**
- **Some additions introduce new control mechanisms serving new goals.**
- **Some additions deal with conflicts between previously existing control mechanisms.**
- **Some changes allow a pre-existing sub-mechanism to be shared (sequentially or concurrently) between a greater variety of other sub-mechanisms in the same architecture.**
- **Some additions allow the processing in one sub-mechanism to be modified dynamically (re-tuned) by another mechanism to meet the latter's needs.**

These are merely a few examples of types of architectural complexification – there are many more.

Architecture determines applicable concepts

Some of the examples illustrate the point that certain descriptions are not applicable to a system if its architecture is too simple.

- **E.g. if the organism O merely had sensors SF and SP, each of which deterministically triggered particular behaviours or randomly selected between two behaviours, then O would have only desire-like (motivational) states, and no belief-like states.**
- **In both examples, desire-like states are examples of affective states insofar as they initiate processes.**
- **If SI is added, and its state-changes can immediately generate or modulate behaviour (e.g. it triggers interventions) then it is not a purely factual sensor: its states are at least in part affective.**
- **The more different kinds of processes make use of SI in the course of making selections or initiating actions, the more the state of SI can be seen as belief-like, rather than affective (desire-like).**

We need to collect many more examples of architectures to see which sorts of concepts they do and do not support.

E.g. none of the architectures just discussed supports the possibility of planning processes, in which sequences of possible actions and their consequences are considered and evaluated, as happens in fully deliberative architectures.

How not to study emotions or most other states of information-processing systems

In physics many things are investigated by studying invariances in their conditional behaviours: input-output contingencies.

E.g. relationships between applied voltage and measured current, or between applied force and measured amount of compression or stretching determine physical properties of objects, like electrical resistance or elasticity.

But this is a bad model to use for information-processing systems, even though many psychologists use it:

- **because any particular collection of observed behaviours can be explained by indefinitely many different information-processing architectures.**
- **because most interesting information-processing systems with multiple interacting concurrent sub-systems capable of learning, development and creative problem-solving will simply not have many invariants of the right sort:**

  as soon as you tell humans about a non-trivial invariant you have discovered they are likely (in many cases) to be able to decide to violate the invariant in future (exceptions are laboratory restricted tasks).

So we need a different approach to categorising types of systems, the states and processes they can be in, and the ways they do things.

We need to consider mechanisms – what sorts of mechanisms?

What sorts of mechanisms?

NOT JUST: matter-manipulating mechanisms

NOT JUST: energy-manipulating mechanisms

BUT MAINLY: information-manipulating mechanisms

But not just computers:

Evolution got there long before we did, so we need to understand something about varieties of organisms.

Among organisms here is enormous diversity, in:

- varieties of behaviours
- diversity of components and sub-functions concurrently or sequentially active
- amount of processing between sensing and acting
- varieties of environments in which they can 'cope' (meet needs)
- time-scales over which individuals and subsystems operate
- requirement for *structural* variation of state (Cf. those with and without 3D vision)
  - Physical systems cannot usually change their physical structure rapidly
  - So evolution produced 'virtual machines' that can, and can have astronomically many distinct structures without changing physical structure
- number and diversity of needs
- potential for conflict between needs and concurrent sub-systems
- methods and mechanisms for detecting and resolving such conflicts
- types of information they can acquire and use about the environment
- types of information they can acquire and use about themselves (as objects in the environment, or as information processors)
- extent of use of *explicit* as opposed to *implicit* information states.
- adaptability
- degree and type of socialisation

We need to understand that diversity

We need a good ontology for varieties of organisms and – organism-like machines.

This will include

- **Kinds of components that can be present**
- **Kinds of functions components can have**
- **Ways of combining components and functions (a grammar for architectures?)**
- **Kinds of interactions there can be between components**
- **Varieties of resulting systems, the behaviours they can produce, the states they can be in, the ways they can change, etc.**

Several different taxonomies are needed, some derived from others

Doing this well may require us to think both about complete systems and their architectures and "sub-atomically" about more basic kinds of building blocks than we normally consider.

Compare how theories about the structure of matter, and processes involving matter have developed.

We need a "sub-atomic theory" of functional roles, parts of which have been sketched above. **(Compare Minsky *The Society of Mind*)**

Towards a general framework

We need to talk about "information-using systems" — where "information" has the everyday sense, not the Shannon technical sense. This notion is being used increasingly in biology.

What are information-using systems?

- **They acquire, store, manipulate, transform, derive, apply information.**
- **The information must be expressed or encoded somehow, e.g. in simple or complex structures – possibly in virtual machines.**
  (The use of *physical* symbol systems is often too restrictive.)
- **These structures may be within the system or in the environment.**
- **The information may be more or less explicit, or implicit.**

A theory of meaning as we normally understand "meaning" in human communication and thinking should be seen as a special case within a general theory of information-using animals and machines.

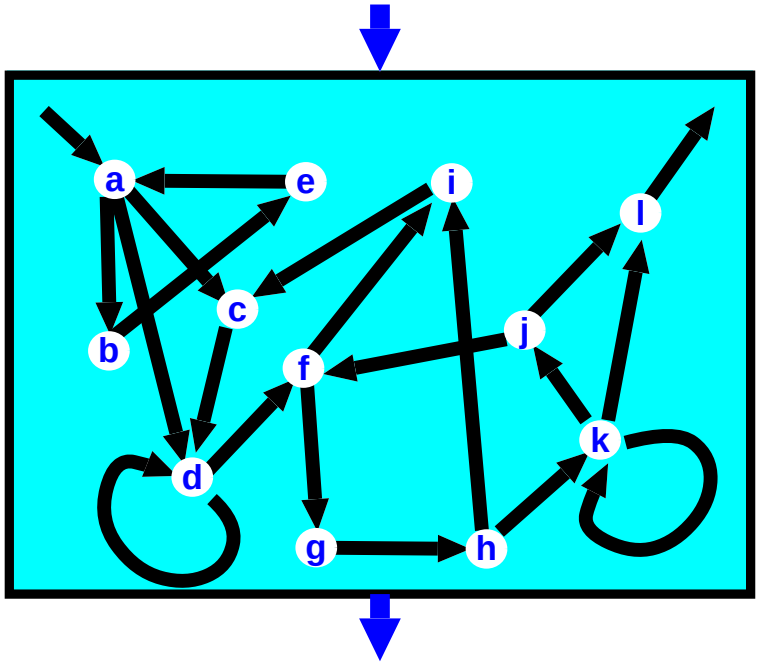
Examples of types of processes involving information

- Acquisition
- Filtering/selecting
- Transforming/interpreting/disambiguating
- Compressing/generalising/abstracting
- Deriving (making inferences, but not only using propositions)
- Storage/Retrieval (many forms: exact, pattern-based, fuzzy)
- Training, adaptation (e.g. modifying weights, inducing rules)
- Constructing (e.g. descriptions of new situations or actions)
- Comparing and describing information (meta-information)
- Reorganising (e.g. formation of new ontologies)
- Testing/interrogating (is X in Y, is A above B, what is the P of Q?)
- Copying/replicating
- Syntactic manipulation of information-bearing structures
- Translating between forms, e.g. propositions, diagrams, weights
- Controlling/triggering/modulating behaviour (internal, external)
- Propagating (e.g. in a semantic net, or neural net)
- Transmitting/communicating
- .... (many more)

The differences involve types of content, types of medium used, and the causal and functional relations between the processes and their precursors and successors.

Functionalism ?

Functionalism is one kind of attempt to understand the notion of virtual machine, in terms of states defined by a state-transition table.



This is how many people think of functionalism: there's a total state which affects input/output contingencies, and each possible state can be defined by how inputs determine next state and outputs.

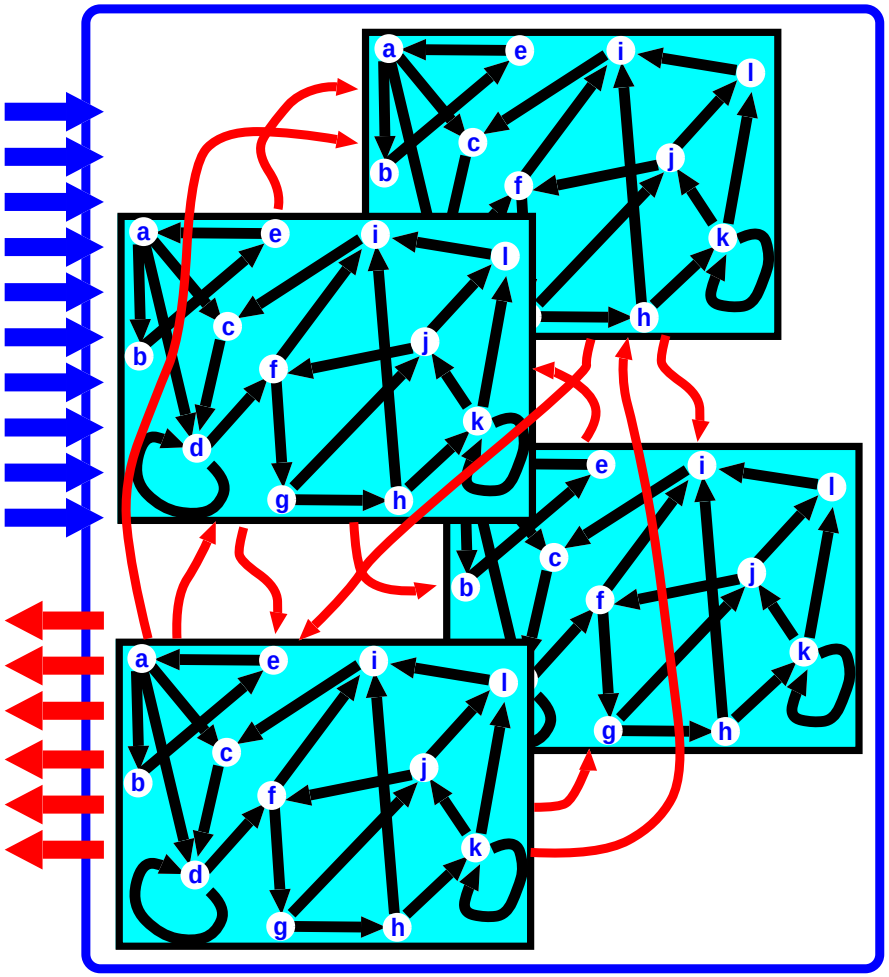
(E.g. see Ned Block's accounts of functionalism.)

HOWEVER THERE'S A RICHER, DEEPER NOTION OF FUNCTIONALISM

Another kind of Functionalism ?

Instead of a single (atomic) state which switches when some input is received, a virtual machine can include many sub-systems with their own states and state transitions going on concurrently, some of them providing inputs to others.

- The different states may change on different time scales: some change very rapidly others very slowly, if at all.
- They can vary in their granularity: some sub-systems may be able to be only in one of a few states, whereas others can switch between vast numbers of possible states (like a computer's virtual memory).
- Some may change continuously, others only in discrete steps.



Some sub-processes may be directly connected to sensors and effectors, whereas others have no direct connections to inputs and outputs and may only be affected very indirectly by sensors or affect motors only very indirectly (if at all!).

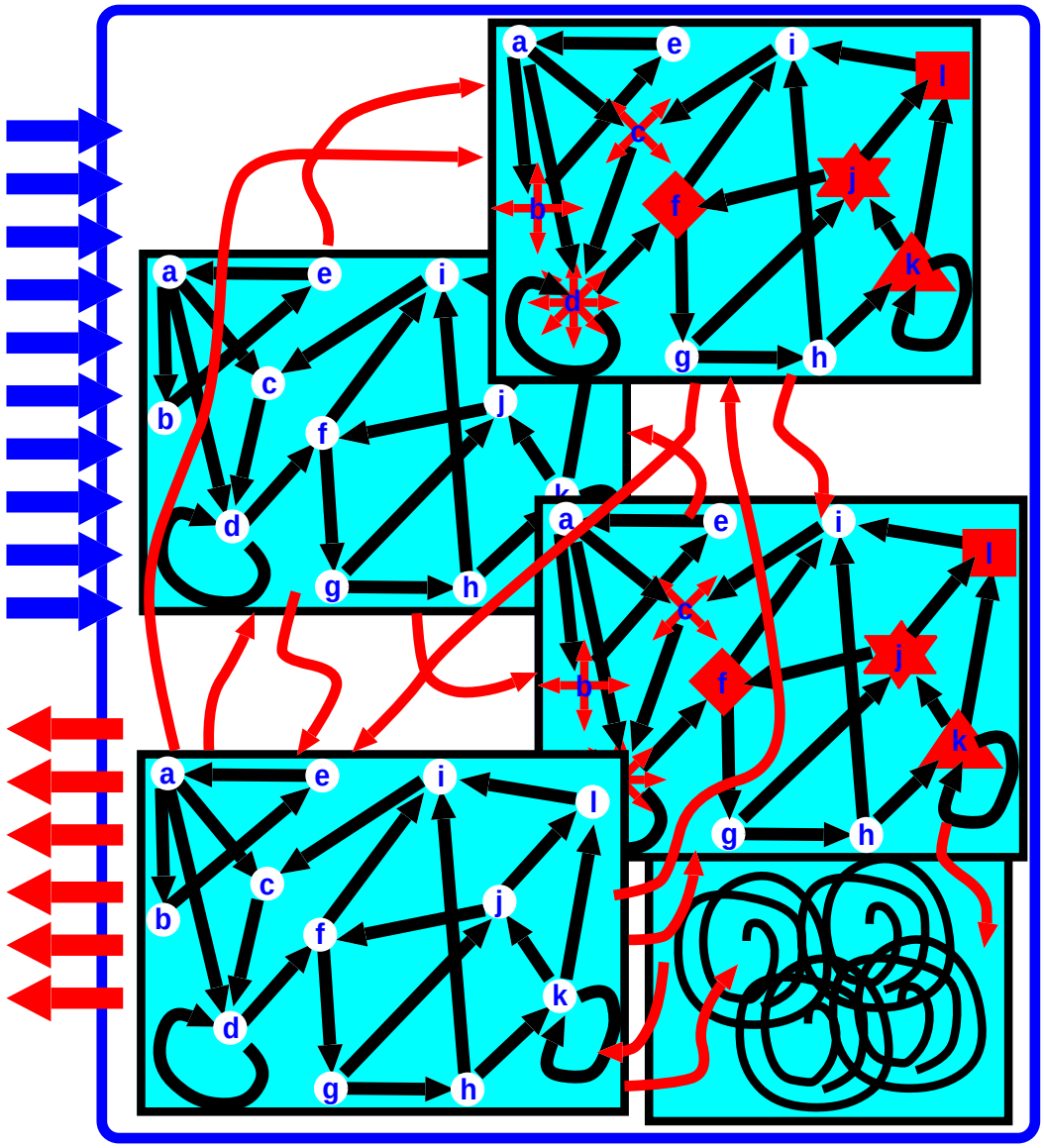
The previous picture is misleading

Because it suggests that the total state is made up of a fixed number of discretely varying sub-states:

We also need to allow systems that can grow structures whose complexity varies over time, as crudely indicated on the right,
e.g. trees, networks, algorithms, plans, thoughts, etc.

And systems that can change continuously, such as many physicists and control engineers have studied for many years, as crudely indicated bottom right
e.g. for controlling movements.

The label 'dynamical system' should be applicable to all these types of sub-system and to complex systems composed of them.



VMF: Virtual Machine Functionalism

We use "Virtual Machine Functionalism" (VMF) to refer to the more general notion of functionalism, in contrast with "Atomic State Functionalism" (ASF) which is generally concerned with finite state machines that have only one state at a time.

VMF allows multiple concurrently active, interactive, sub-states changing on different time scales (some continuously) with varying complexity.

VMF also allows that the Input/Output bandwidth of the system with multiple interacting internal states may be too low to reveal everything going on internally.

There may still be real, causally efficacious, internal virtual machine events and processes that cannot be directly observed and whose effects may not even be indirectly manifested externally.

Even opening up the system may not make it easy to observe the VM events and processes (decompiling can be too hard).

If some links between systems can be turned on and off by internal processes, then during some states:

some of the sub-systems may not have any causal influence on outputs.

Those running sub-systems still exist and can include internal causal interactions within and between themselves: scientific investigations will have to allow for this possibility.

See also http://www.cs.bham.ac.uk/research/cogaff/talks/#inf

NOTES

- The previous points all need to be developed in more detail and with more precision.
- A machine or organism may do some of these things internally, some externally, and some in cooperation with others: information-processing need not be internal. (The same calculation can be done in your head or in sand.)
- The processes may be discrete or continuous (digital or analog).
- Some people think information is inherently static and incapable of causing processes to occur. They forget the reasons why we say things like:
  - "The pen is mightier than the sword"
  - "News about Diana's death caused expressions of grief in many countries."
  - "His refusal made me very angry."
  - "The corporal's command made the men jump to attention".

Information has causal powers when it enters, or is created in, a situation where it can initiate a new process or modulate an old one: like dropping a crystal into a super-cooled liquid.

Many such situations are familiar: news can be a "bombshell". An idea can make you turn around and go home. A syntax error in a program can cause compilation to be aborted.

It is important not to forget that there's such a thing as control information.

This is why such things as desires, moods and emotions can be accommodated within an information-processing theory of mind.

Requirements for Information-Processing

Not all the processes listed previously are possible in all architectures.

E.g. constructing and comparing descriptions of possible future actions, needs a "workspace" for items of varying complexity.

Some kinds of neural net require mechanisms supporting continuous variation.

Some kinds of manipulation require an engine able to construct and manipulate "Fregean" structures, with hierarchic function plus arguments decomposition. (E.g. f(g(a, h(b,c)), h(d,e)))

We must distinguish requirements specified (a) in terms of a virtual machine architecture (b) in terms of physical mechanisms.

A VM SPECIFICATION might mention a strict stack discipline for procedure activations, with local variables and return address in each stack frame.

A PHYSICAL SPECIFICATION might mention fast special purpose registers, etc.

How much the properties of a particular VM can be decoupled from properties of the physical implementation will vary.

How much of a VM is implemented in the "external" environment will vary. (E.g. pheromone trails used by insects.)

METHODOLOGICAL POINT

We cannot expect deep understanding
if we merely develop one architecture without exploring:

- **Design space: the space of possible architectures**
  (or at least a relevant neighbourhood in that space)

  Compare John Barrow's book *Constants of Nature*
  (exploring alternative versions of general physical/cosmological theories).

- **Niche space: the space of possible sets of requirements for architectures**
  (or at least a relevant neighbourhood in that space)

- **Relationships between (the relevant neighbourhoods in) design space and niche space.**

However we can sometimes solve practical problems without deep science. **(See http://www.cs.bham.ac.uk/research/cogaff/talks/#talk4)**

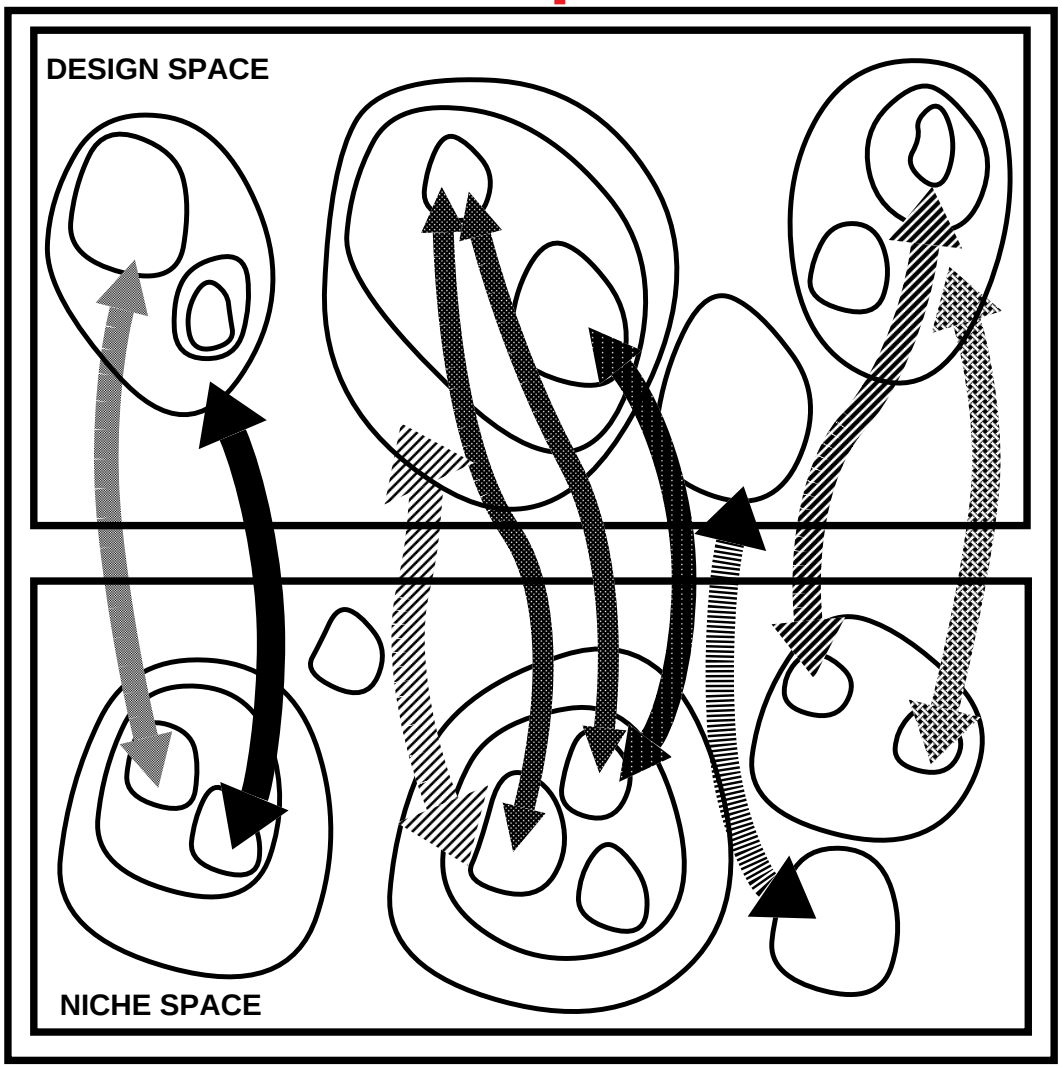
Design space and niche space

Design space: the space of possible architectures.

Niche space: the space of possible sets of requirements for architectures.

There are discontinuities in both design space and niche space: not all changes are continuous (smooth).

Do not expect one fitness function.
Instead expect *diverse structured fitness relations* between designs and niches.

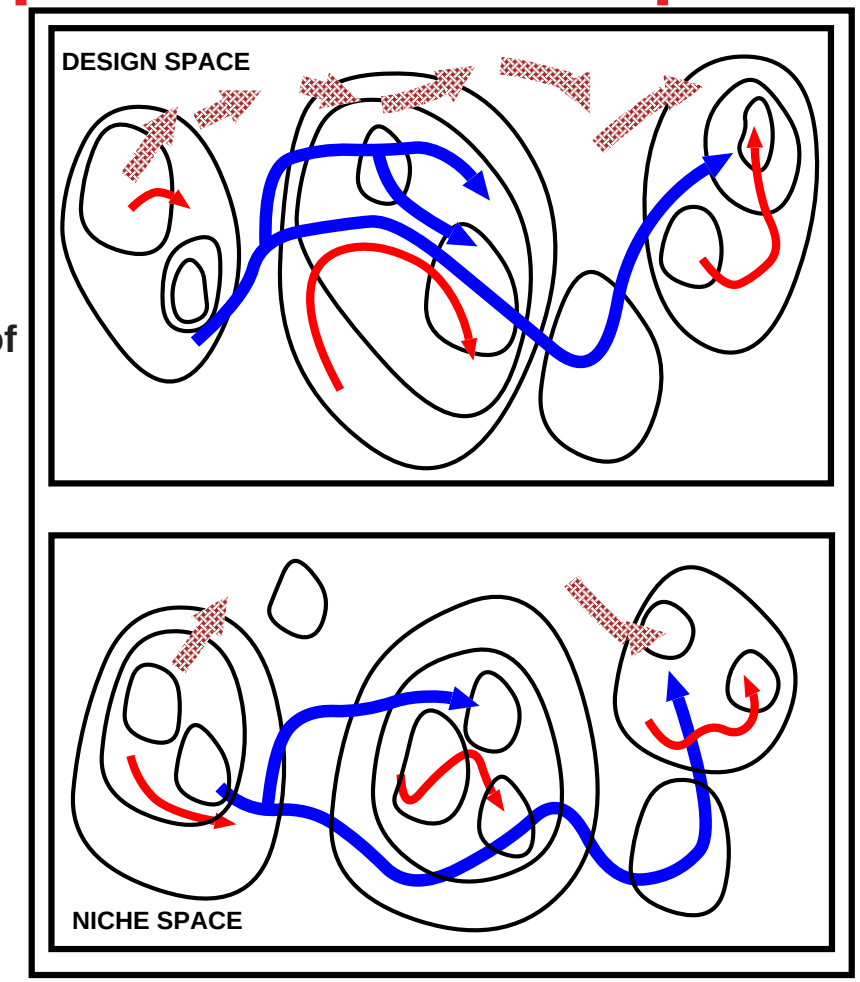


Trajectories in design space and niche space

There are different sorts of trajectories in both spaces:

- **i-trajectories:**
  Individual learning and development
- **e-trajectories:**
  Evolutionary development, across generations, of a species.
- **r-trajectories:**
  Repair trajectories: an external agent replaces, repairs or adds some new feature. The process may temporarily disable the thing being repaired or modified. It may then jump to a new part of design space and niche space.
- **s-trajectories:**
  Trajectories of social systems.

Some e-trajectories may be influenced by cognitive processes (e.g. mate-selection). We can call them c-trajectories (not shown separately).



Different motivations for interest in implementable models of emotions

(i) Science and philosophy:

An interest in natural emotions (in humans and other animals) as something to be modelled and explained, or an investigation of how they might have evolved, etc.

(ii) Improved interaction:

A desire to give machines which have to interact with humans an understanding of emotions as a requirement for some aspects of that task (Sloman 1992)

(iii) Entertainment:

A desire to produce new kinds of computer-based entertainments where synthetic agents, e.g. software agents or "toy" robots, produce convincing emotional behaviour.

(iv) Education:

Using models of type (i), (ii), (iii) etc. in educational tools for trainee psychologists, therapists, etc.

(v) Therapy, counselling, etc.:

If we have a better understanding of the nature of the emotions and other affective states, and their architectural underpinnings, we may be better able to provide helpful therapy when needed.

The conceptual requirements for these objectives are different.

E.g. “believable” behaviour in constrained contexts could be the product of widely different models, including at one extreme very large, hand-coded lookup tables specifying what to do when.

But in the long run a deep and accurate model of the first type may be required for effectively achieving even the engineering goals of types (ii) and (iii)

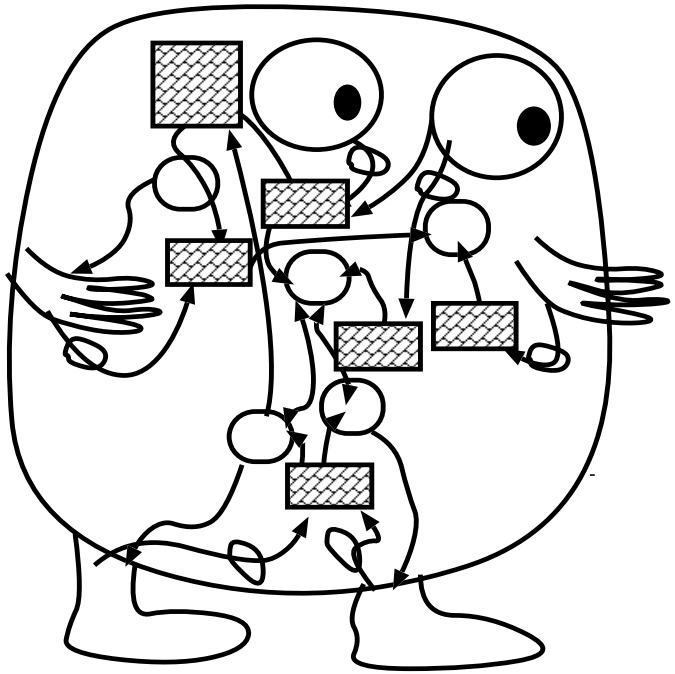
For now we address only goal (i) (Science and philosophy, including conceptual analysis), while keeping an eye on the requirements for the others.

NOTE: I am not specially interested in *emotions* except as a special case of a wide range of phenomena that need to be accommodated in a theory of what minds are (an ontology for minds) and explanations of how they work.

Here is a very simple toy demo of type (iii)/(iv).

(Show sim_feelings demo
http://www.cs.bham.ac.uk/research/poplog/sim/teach/sim_feelings)

WHAT SORT OF ARCHITECTURE?
Could it be an unintelligible mess?



YES, IN PRINCIPLE.
As some have argued.

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

Problem 1:

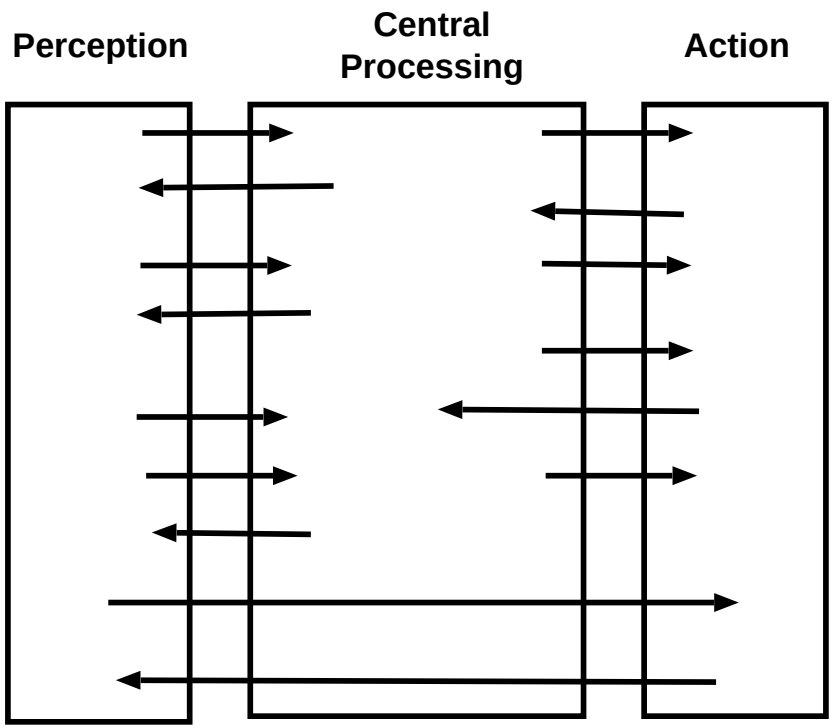
Time required and variety of contexts required for a suitably general design to evolve.

Problem 2:

Storage space required to encode all possibly relevant behaviours if there's no "run-time synthesis" module.

Towards a unifying theory of architectures for natural and artificial agents

1. A "triple tower" perspective



There are many variants,
e.g. Nilsson, Albus....

Systems can be
"nearly decomposable".
(Herbert Simon)

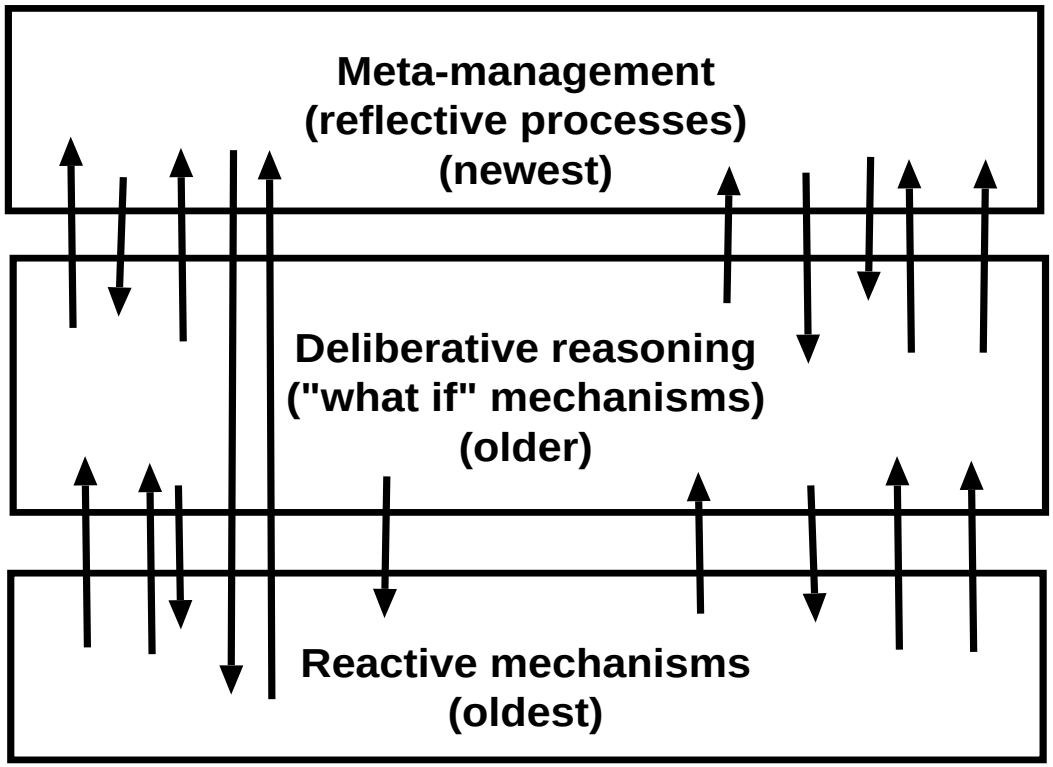
Boundaries can change with learning and development.

Many theories fail to do justice to the complexities of perception and action, because people do not analyse the requirements in depth.

There can be *hierarchies* of percepts and actions.
See http://www.cs.bham.ac.uk/research/cogaff/talks/#talk21

ANOTHER COMMON ARCHITECTURAL PARTITION (functional, evolutionary)

2. A “triple layer” perspective



(MANY VARIANTS – FOR EACH LAYER)

This is not the “three layer”system of Eran Gatt.

Features of the layers

- **Reactive** systems can be highly parallel, very fast, and use a mixture of analog circuits and digital, e.g. rules.
  - Some reactive capabilities may be innate, others learnt.
  - Reflexes, with direct connections from sensors to motors, could be separated out from the other reactive mechanisms.
- **Deliberative mechanisms** are inherently slow, serial, knowledge-based, resource limited.

  Sophisticated deliberative systems, with powerful formalisms for expressing descriptions of alternative possibilities, require a lot of supporting mechanisms, which may not evolve often, because of their cost, e.g. requiring "expensive" brain mechanisms (at peak of food pyramid)
- **Meta-management** uses additional mechanisms for monitoring, categorising, evaluating, and in some cases modifying or controlling internal states and processes.

  In sophisticated organisms meta-management (and other layers) may use culturally determined categories and procedures (e.g. in guilt and self-torment.)

The layers may be concurrent or interleaved or pipelined.
There may or may not be a dominance hierarchy.

Layered architectures have many variants

With different subdivisions and interpretations of subdivisions, and different patterns of control and information flow.

Different principles of subdivision in layered architectures

- evolutionary stages
- levels of abstraction,
- forms of representation used
- kinds of mechanisms used
- which bits use analog mechanisms, which digital
- how much concurrency
- whether components are synchronised
- how many concurrent goals can exist
- how goal conflicts are resolved (arbitration, preferences....)
- control-hierarchy,
  (Top-down vs multi-directional control. See subsumption architectures, below.)
- information flow
  (e.g. the popular 'Omega' Ω model of information flow, described below.)

Beware of terminological differences

E.g.

- **Some people use "reactive" to mean stateless – we don't.**
- **Some people use "reactive" to exclude chained automatic routines – we don't.**
- **Some people use "reactive" to exclude the use of discrete symbols – we don't.**
- **Some people use "reflective" to refer to "watchful" processes of plan-execution leading to plan improvements, etc. – we include that under "deliberative."**
- **Our top layer ("meta-management") requires special architectural support for an internal process observing, categorising, evaluating, and possibly controlling or modulating other internal processes.**
  Perhaps it could evolve by copying and redeploying "alarm" mechanisms.

COMBINING THE VIEWS: LAYERS + PILLARS = GRID

An architectural “schema” (CogAff) not an architecture.

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

On some versions of this diagram I add lots of arrows between boxes indicating possible routes for flow of information (including control signals) – in principle, any two boxes can be connected in either direction.

Not all organisms will have all the kinds of components, or connections.

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

SENSING AND ACTING CAN BE ARBITRARILY SOPHISTICATED

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

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

  Think of the difference between
  - perceiving edges, optical flow, texture gradients
  - perceiving chairs, tables, support relations
  - perceiving happiness, surprise, anger, which way someone is looking.

  We understand very little about what affordances are, how they are represented, how they are perceived, how they are used.

  A draft paper is here:

  http://www.cs.bham.ac.uk/research/cogaff/sloman-diag03.pdf

Perception goes far beyond segmenting, recognising, describing what is “out there”

It includes:

- providing information about *affordances*
  (Gibson was closer to the truth than Marr)
- directly triggering physiological reactions
  e.g. posture control, sexual responses)
- evaluating what is detected,
- triggering new motivations
- triggering “alarm” mechanisms
- . . . . .

AND THESE ALL NEED INTERNAL LANGUAGES OF SOME SORT

Multi-window perception

We'll propose a "multi-window" theory of vision (and action, and possibly other modalities of perception.)

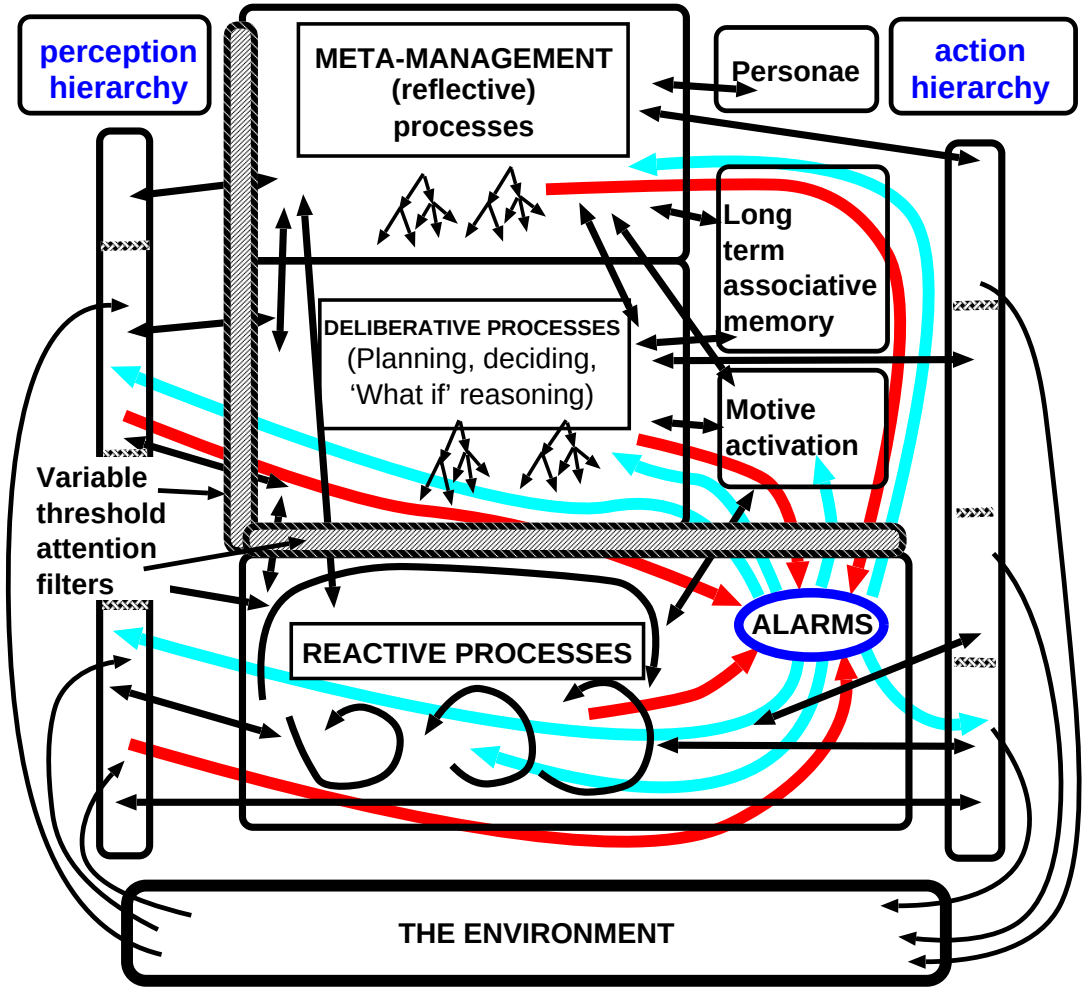
- The key idea is that visual mechanisms operate on different levels of abstraction in parallel, combining bottom up, top down, and background information in very flexible ways.
- The different levels may use different ontologies to characterise what is perceived.
- Some of them can be concerned with evaluation as well as description.
  (Compare L.Pryor's "reference features")
- Contrast "peephole" perception: sensory information is a homogeneous collection of information processed in homogeneous ways (e.g. statistical methods).

Architectural theories which ignore the possibilities of multi-window perception and action will fail to account for some of the complexities of human minds.

This may not matter for some engineering applications.

The architecture of a human mind

(very sketchy first draft – see http://www.cs.bham.ac.uk/research/cogaff/)



See other cogaff papers and talks for details

THE BIRMINGHAM COGNITION AND AFFECT PROJECT

OVERVIEW:

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

PAPERS:

http://www.cs.bham.ac.uk/research/cogaff/

(References to other work can be found in papers in this directory)

TOOLS:

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

http://www.cs.bham.ac.uk/~axs/cogaff/simagent.html

(the SIM_AGENT toolkit)

DEMO-MOVIES:

http://www.cs.bham.ac.uk/research/poplog/figs/simagent/

SLIDES FOR TALKS:

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