Slides Prepared for Talk At
The Voice Box
Royal Festival Hall
London
Saturday 21 Feb 1997

http://www.sbc.org.uk/literate.htm

WHAT SORTS OF MACHINES CAN LOVE?

Architectural Requirements

for

Human-like Agents

Both Natural and Artificial.

Aaron Sloman
School of Computer Science

Cognitive Science Research Centre

The University of Birmingham, UK

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

ASK THE GHOST OF GILBERT RYLE

Gilbert Ryle's 1949 book THE CONCEPT OF MIND **has many important and relevant ideas. But he lacked the conceptual tools required for expressing the ideas, and as a result was wrongly interpreted as a behaviourist.**

HOW, NOT WHETHER

Can machines have emotions?

We know that machines can have emotions,
for humans are machines,
though not in the sense of being artefacts.

WE ARE BIOLOGICAL ORGANISMS:

- **rooted in carbon, hydrogen, oxygen, nitrogen, iron and other physical stuff,**
- **evolved through millions of years of exploration,**
- **revealing our history in our design,**
- **grown in wombs, cots, playgrounds, and cultures,**
- **acquiring and discarding energy and matter,**
- **acquiring, transforming and storing information,**
- **writing poems, plays and newspaper reports,**
- **providing the stuff to write about,**
- **deceiving ourselves that we are different,**
- **wanting the truth to be** THUS **…,**
  rather than wanting to know the truth.

BUT WE ARE NOT "JUST" MACHINES, "MERE" MACHINES

ANY MORE THAN COMPUTING SYSTEMS ARE.
(Beware the "nothing buttery" fallacy.)

WE, AND THEY, ARE MULTI-LEVEL MACHINES:

We have physical architectures and we have information processing architectures.

OUR INFORMATION PROCESSING INCLUDES:

- **generating goals**
- **considering options**
- **making selections**

JUST LIKE DEEP BLUE

AND ALSO

- **detecting our own states**
- **evaluating our own thoughts and reasons**
- **feeling ashamed, guilty, fearful, excited, self-satisfied, infatuated**

JUST LIKE AI SYSTEMS OF THE (DISTANT?) FUTURE

KEY NOTIONS:

1. **Architecture**
2. **Virtual machines**

WHAT SORT OF ARCHITECTURE COULD SUPPORT BEING IN LOVE?

One requirement:

X is in love with Y IMPLIES

X's thoughts are constantly drawn to Y.

The hard part is understanding the required information processing architecture:

Providing the body parts is not enough, and may not even be necessary. (One kind of love...)

It may help if we relate this to some familiar types of things, with different architectures.

- **Could an ant be in love?**
- **Is a drake in love with its duck?**
- **Could a chimp be in love?**
- **Could a new born baby be in love?**

NOTE:
Being in love is not the same sort of thing as loving.
Love in general is an attitude and need not be emotional:

- **you can love members of your family without constantly dwelling on them**

YOU CAN

- **love your country**
- **love the organisation you work for**
- **love football ...**

ARCHITECTURAL LAYERS

An architecture explains what sorts of states and processes are possible.

THE LARGER TASK :

To devise a theory of possible types of architectures and use the architectures as frameworks for generating families of descriptive and explanatory concepts applicable to:

- **different sorts of humans (including infants and people with various kinds of brain damage)**
- **different sorts of animals**
- **different sorts of artificial agents.**

CONJECTURE:

1. Human-like agents need an architecture with at least three layers.

- **A very old reactive layer, found in various forms in all animals, including insects).**
- **More recently evolved deliberative layer, found in varying degrees of sophistication other animals.**
- **An even more recent meta-management (reflective) layer providing self-monitoring and self-control, perhaps found in simple forms only in other primates. (Probably not in very young children?)**

2. These three layers are found both in the central information processing system and also in other components, e.g. sensory systems, action systems.

3. Additional modules support or modulate these functions, e.g.

- **A global "alarm" system (the limbic system?)**
- **Various kinds of associative information stores (long term memory, essential for "what if..." deliberations.)**
- **Motive generation, comparison, selection modules**
- **....other things....**

DIFFERENT LAYERS EXPLAIN DIFFERENT SORTS OF EMOTIONAL PROCESSES

For example:

(1) emotional states (like being startled, terrified, sexually stimulated) based on the old reactive layer and global alarm system shared with many other animals,

(2) emotional states (like being anxious, apprehensive, relieved, pleasantly surprised) which depend on the existence of the deliberative layer, in which plans can be created and executed, risks assessed, success detected, etc.

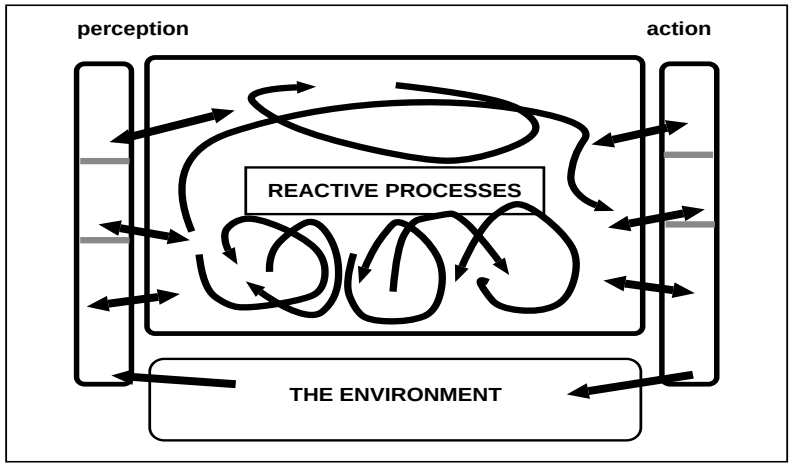
(3) emotional states (like feeling humiliated, infatuated, guilty, or full of excited anticipation) in which attempts to focus attention on urgent or important tasks can be difficult or impossible, because of processes involving the meta-management layer being interrupted or diverted by other processes (in the reactive layer or global alarm system).

This framework eliminates a considerable amount of argumentation at cross-purposes, e.g. about the nature of emotions, because people are talking about different sorts of things without a theoretical framework in which to discuss the differences.

(THERE ARE DOZENS OF DIFFERENT DEFINITIONS OF "EMOTION" IN PSYCHOLOGICAL AND PHILOSOPHICAL LITERATURE.)

REACTIVE AGENTS

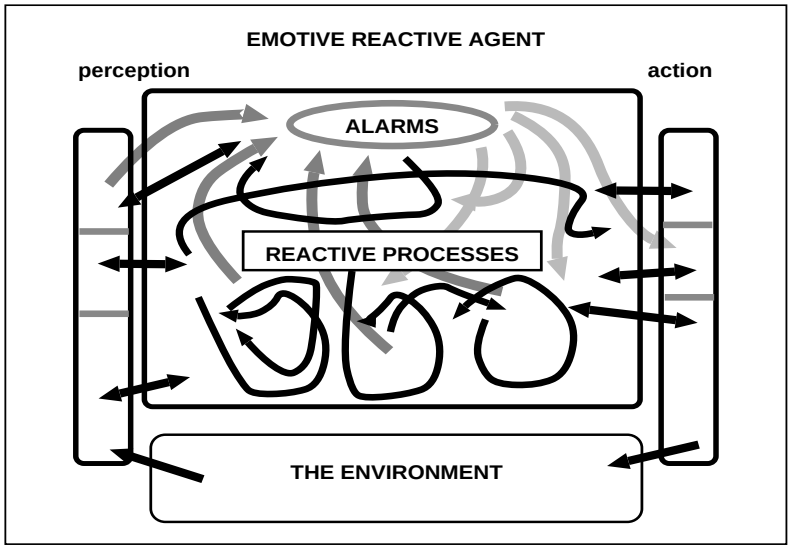
HOW TO DESIGN AN INSECT?



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 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**
- **The agent can survive even if it has only genetically determined behaviours**
- **Difficulties arise if the environment requires new plan structures.**
- **This may not matter if individuals are cheap and expendable (insects?).**

EMOTIVE REACTIVE AGENTS

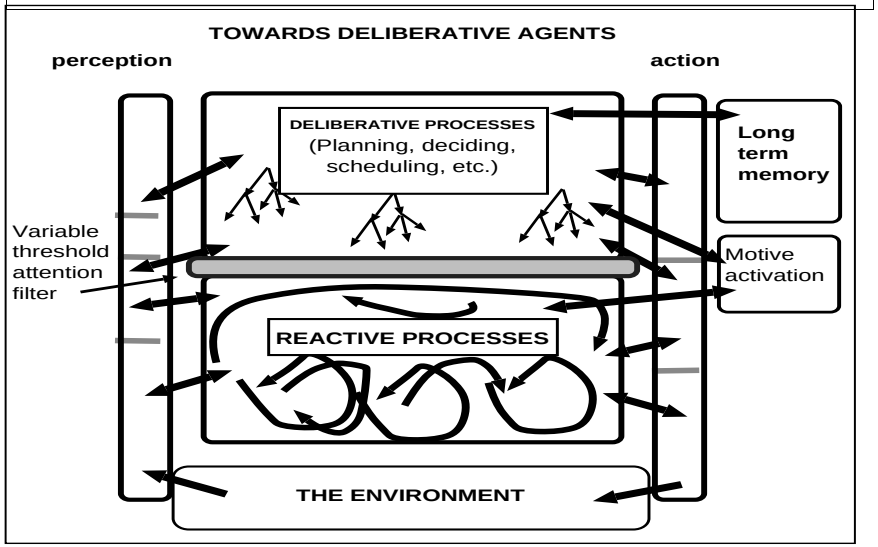


Some sort of "override" mechanism seems to be needed for certain contexts

AN ALARM MECHANISM:

- **Allows rapid redirection of the whole system**
- **sudden dangers**
- **sudden opportunities**
- FREEZING
- FIGHTING
- FEEDING
- ATTENDING
- FLEEING
- MATING
- MORE SPECIFIC TRAINED AND INNATE AUTOMATIC RESPONSES

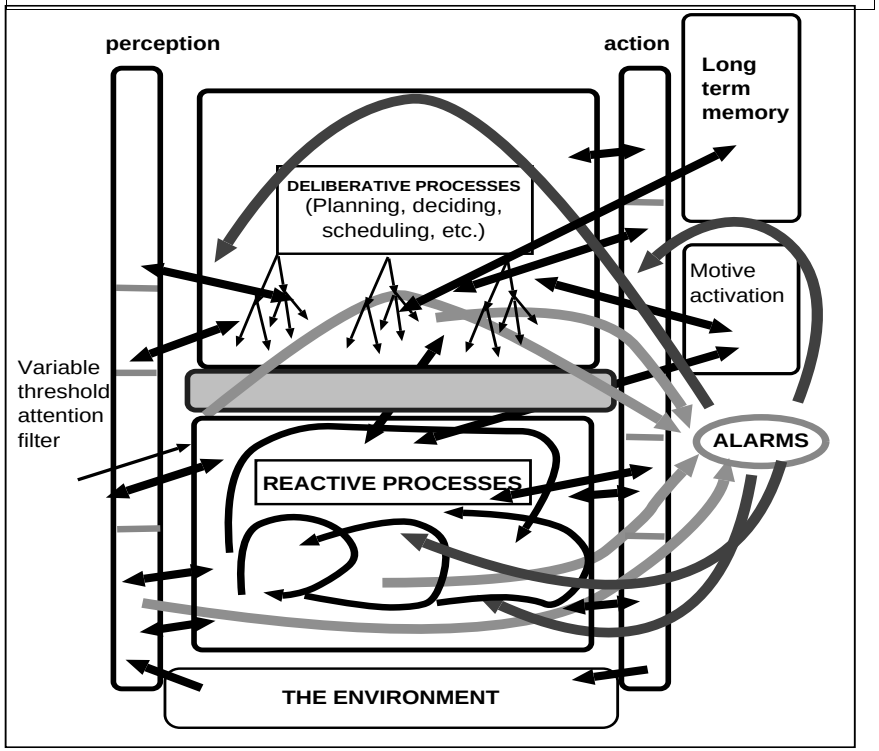
REACTIVE AND DELIBERATIVE LAYERS



IN A DELIBERATIVE MECHANISM:

- **Motives are explicit and plans are 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**
- **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**
- **Mating responses**
- **Specialised learnt responses**

On some viewers the textured arrows into the “alarms” box will not show up on the diagram

SELF-MONITORING (META-MANAGEMENT)

Deliberative mechanisms with evolutionarily determined strategies may be too rigid.

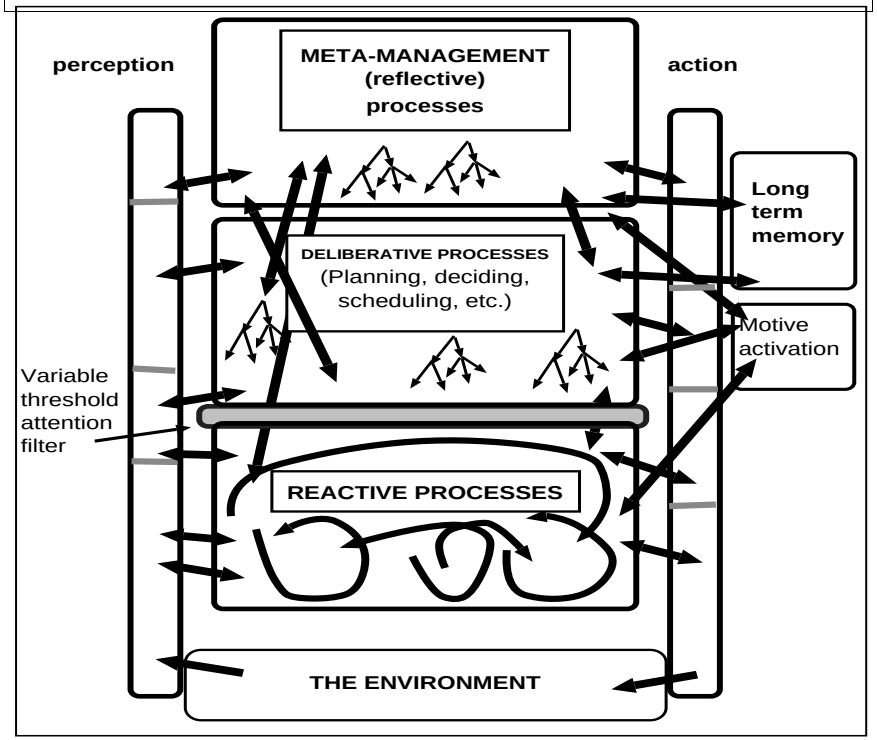
Internal monitoring mechanisms may help to overcome this if they

- **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.**

META-META-MANAGEMENT MAY NOT BE NEEDED

IF META-MANAGEMENT MECHANISMS ARE RECURSIVE!

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?)**

"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.**

ALARM MECHANISM CAN BE EXTENDED

(I.e. the limbic system??)

- **Inputs from all parts of the system**
- **Outputs to all parts of the system**
- **Fast (stupid) reactions**

(Too complex to add to diagram: imagine 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.)

NOTE: In humans 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.). Replication of their functions using only computing mechanisms may be difficult, though not necessarily impossible.

ARCHITECTURE AND EMOTION

Different architectural layers support different sorts of emotions:

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

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

  etc. etc.

The DELIBERATIVE **layer enables:**

- **being anxious or apprehensive about things going wrong**
- **being frustrated by failure**
- **excitement at anticipated success**
- **being relieved at avoiding danger**
- **being relieved or pleasantly surprised by success**

  etc. etc.

The SELF MONITORING META-MANAGEMENT **layer, explains:**

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

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

BEING IN LOVE

This involves very complex interactions between many parts of the system, including information which is "dormant" much of the time.

The phenomenon cannot be explained, except as part of a wide ranging explanatory architecture.

Read what poets and novelists and playwrights say about love, and ask yourself: what kinds of information processing mechanisms are presupposed.

Shakespeare wrote:

LOVE IS NOT LOVE WHICH ALTERS WHEN IT ALTERATION FINDS

This implies that lovers can find alteration, i.e. perceive things, including perceiving changes. Love is one of those states which can change, and be influenced by new percepts, new discoveries. However, the claim is that it is not one which is easily changed, e.g. by detecting a change in the beloved.

Compare Sir John Suckling

OUT UPON IT. I HAVE LOVED,
THREE WHOLE DAYS TOGETHER
AND AM LIKE TO LOVE THREE MORE,
IF IT PROVE FAIR WEATHER

And then there's the chemical infrastructure:

Calverly

THE HEART WHICH GRIEF HATH CANKERED
HATH ONE UNFAILING REMEDY — THE TANKARD

Extracts from Oxford dictionary of quotations

MORE ON META-MANAGEMENT

The global "ALARM" mechanism can have inputs and outputs linked to all three layers.

Learning to control or suppress them is part of emotional maturity.

The meta-management architectural layer may be "occupied" by different "control regimes" at different times:

- **Behaving with your family**
- **Driving a car**
- **Being a ruthless manager at work**

A basis for understanding multiple personality disorders? What are the "role-switching" mechanisms? How can they go wrong? How can abuse in infancy produce long term damage in the architecture?

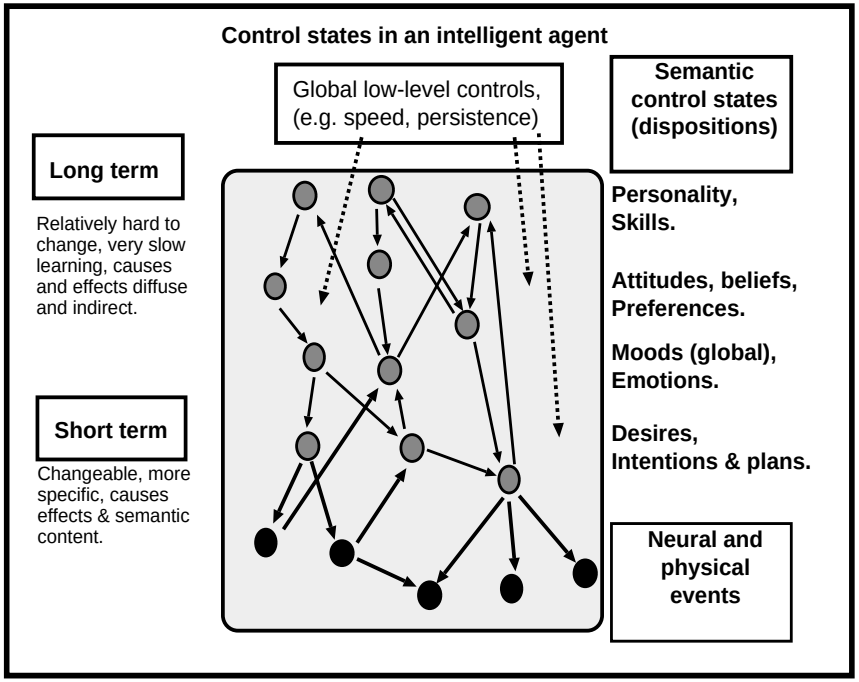
SOME FEATURES ARE "EMERGENT" NOT DIRECTLY IMPLMENTED

NOT EVERYTHING SUPPORTED BY A MECHANISM IS PART OF ITS FUNCTION:

MULTI-PROCESSING OPERATING SYSTEMS SUPPORT THRASHING, BUT DON'T HAVE A TRHASHING MECHANISM!

SOME FUNCTIONAL MECHANISMS HAVE DYSFUNCTIONAL CONSEQUENCES.

TYPES OF CONTROL STATES



Control states of varying scope and duration

The "higher" states are:

- **Harder to change**
- **More long lasting**
- **Subject to more influences**
- **More general in their effects**
- **More indirect in their effects**
- **More likely to be genetically determined(??)**

Different control states have different underlying mechanisms. E.g. some of these are likely to be chemical, others more concerned with procedural and declarative information structures.

FORMS OF LEARNING & DEVELOPMENT

When there is such a complex architecture there are many different forms of development or learning

- **Modification of weights in a reactive subsystem (e.g. produced by reinforcement learning)**
- **Creation of new links between systems**
- **Creation of new subsystems (??)**
- **Creation of new forms of representation**
- **Storage of new information (particular, general)**
- **Development of new "chains" in a reactive mechanism**
- **Development of new plans, stored for future use**
- **Transfer of slow serial processes from deliberative to reactive subsystem (to the cerebellum?), forming new fluent skills.**
- **Development of new links (triggers) connected to such new "skils"**
- **Development of new forms of motivation, new motive generators, new motive comparators**
- **Development of new filtering strategies**
- **Development of new meta-management strategies, e.g. attention control strategies**
- **Learning new ways to control/suppress the effects of the "alarms" submechanism. (E.g. ignoring those which turn out to be misleading.)**

ALSO: many ways in which damage, disease, or genetic disability can interfere with normal functioning.

THERE IS NO UNIQUE ARCHITECTURE

Many architectures are needed for different organisms or artificial agents.

Even humans differ from one another: children, adolescents, adults and senile adults.

Naturally occurring alien intelligences and artificial human-like agents may turn out to have architectures that are not exactly like those of normal adult humans.

Different architectures support different classes of mental states.

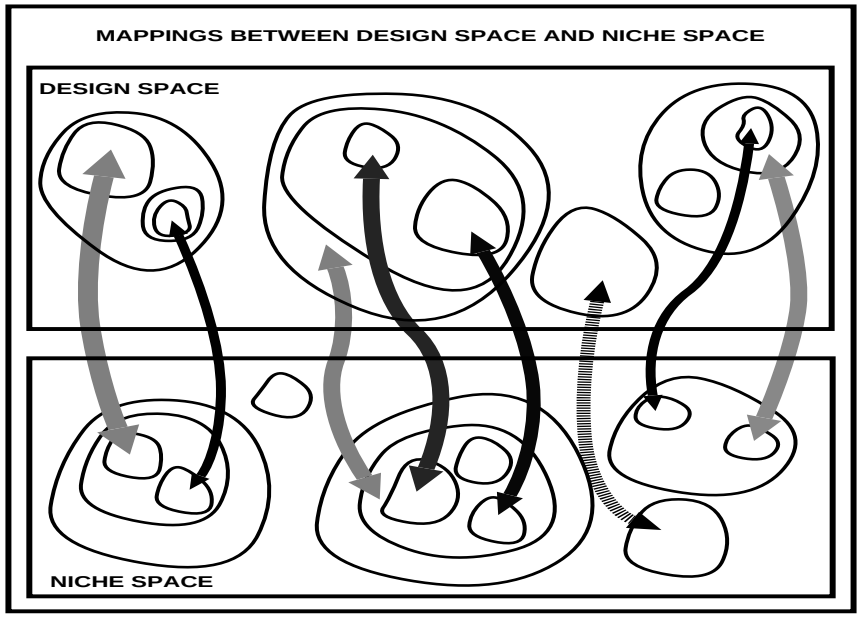
Designers of synthetic agents need to be aware of the evolutionary pressures behind human architectures. Some artificial agents may need similar architectures.

There may be some unanticipated consequences of these design features (SLOMAN AND CROUCHER IJCAI 1981)**.**

Analysing these possibilities is hard.

So we need to explore relationships between "niche space" and "design space".

DESIGN SPACE and NICHE SPACE



- **A niche is a set of requirements**
- **A design is a set of specifications**
- **Mappings are not unique: trade-offs everywhere**
- **Designs need no designer, requirements no requirer.**

DYNAMICS – Which trajectories are possible:

- **Within an agent (development, learning, falling in love)?**
- **Across generations (evolution, ALIFE)?**
- **Only based on external manipulation?**

We need to ask not only which sorts of architectures can support love, but also how they can develop and change, and how they might have evolved. This is part of the study of trajectories in niche space and design space.

NOTE The "Turing test" defines a tiny niche region ....

of relatively little interest, except as a technical challenge.

All this is a topic for another (long) talk.