# Towards a long term AI project, re-combining various fragmented areas of AI (and other disciplines)

**These slides represent my on-going attempt to understand and summarise some of the implications of some recent discussions on how to design an ambitious AI project to help re-vitalise and re-integrate AI research on "complete" minds.**

**This is not intended as a final, polished report or proposal.**

**The proposal is biased by my interests and the directions adopted in the Birmingham Cognition and Affect project over the last 11 years or so, though the work has had many influences outside Birmingham especially the writings of John McCarthy, Marvin Minsky and Herbert Simon (in alphabetical order!)**

**These notes constitute an invitation to comment on and criticise the ideas expressed herein.**

**And an invitation to collaborate on the (very difficult) tasks described.**

---

**Last changed: November 10, 2002.**
**Aaron Sloman http://www.cs.bham.ac.uk/~axs/**
**For some background ideas, see also http://www.cs.bham.ac.uk/research/cogaff/ibm02/**
**Comments will be posted here http://www.cs.bham.ac.uk/research/cogaff/manip/comments.html**

# Acknowledgements

These sketchy, incomplete slides present some thoughts about a possible project, partly stimulated by discussions at the Workshop on 14-16th April 2002, in St Thomas, Virgin Islands, led by Marvin Minsky and Push Singh.

The slides owe much to suggestions made at the workshop by Marvin Minsky and Push Singh (MIT).

See Minsky's draft chapters for The Emotion Machine
http://www.media.mit.edu/~minsky/

Also McCarthy on designing a child:
http://www-formal.stanford.edu/jmc/child1.html

Some of the background to this proposal can be found in papers available at the CogAff web site: http://www.cs.bham.ac.uk/research/cogaff/

Thanks also to Ken Forbus for helpful comments and criticisms in response to an early draft.

I've probably unwittingly used ideas presented by others at the workshop.

There is also overlap with at least the objectives of the "Cog" project led by Rodney Brooks at MIT, though many of the emphases here are different.

# Towards a project specification

- **The first major milestone of the project will be to produce two interacting child-like robots in a world of manipulable objects**
  **(requiring several years of multidisciplinary research and development)**
- **Initially the robots will be simulated, while we get a feel for some of the high level design issues. Later on, physical versions may be built.**
  **(However, some collaborators may wish to start with physical robots.)**
- **The robots will be as much as possible (which at first will not be very much) like a pair of children aged between 3 and 5, one of whom is old enough to be able to explain things to the other, or teach the other, either when asked for help or merely when the need for help is evident.**
- **The robots will interact with each other and with toys and other objects in various parts of the house. They should also be able to interact with the researchers, with other adults, and possibly with real children who are old enough to be able to cope with the relative stupidity and artificiality of the robots. (Interaction could be via text, mouse, and later speech input.)**
- **Much later (after many years) the robots should be able to interact with young human children as equals.**
- **For much of the project the two robots will model idealised and simplified children.**
  **(In the near future there is no hope of mirroring all of the physical structure and brain function of a real child. Like Galileo thinking about motion we shall have to choose our simplifications very carefully so as to address deep challenges without getting bogged down in side-issues.)**

# Why do this?

- One reason is that the attempt to model various capabilities of young children is of intrinsic interest.
- Another reason is the 'hunch' that by the age of about five years, normal children have acquired a huge amount of skill and understanding required for interacting with many objects in the physical environment, with other people, and with themselves (e.g. finding answers to their own questions).
- This highly integrated, deep and diverse collection of capabilities seems to provide the 'infrastructure' for many kinds of knowledge and skills acquired later, including linguistic skills, social skills and many kinds of things taught in school or available in books etc.
- If so, a wide variety of types of robots could be produced by giving appropriate forms of teaching to the initial robots.
- Another possibility worth exploring is that a humanoid with something like a five year old's capabilities can be designed with interfaces enabling a wide variety of 'plugin' packages to provide new kinds of knowledge and skills.
- It is also conceivable that our research will help us understand why that is impossible, i.e. why something like human education is needed to teach the robot adult English and other languages, mathematics, geography, history, biology, medicine, law, etc.

# Why not start with a new-born infant?

## There are many possible earlier stages than a four or five year old: would it not be easier to start with some earlier stage and teach the infant as we teach a human infant, or even start with a simulated embryo?

- I believe this would be much harder to do for various reasons including the difficulty of finding out what the capabilities are of a much younger child with whom communication is very difficult, as any parent knows. As language develops it provides a powerful additional window into (and out of) the mind.
- We cannot hope in an AI project to solve the problems of embryology.
- It is possible that by developing a deep model of a five year old we shall be in a better position to understand requirements for a developing and learning system that starts at a much earlier stage of development.
- Finally, even if we could simulate the learning and development by training an artificial infant we might end up understanding no more about the result of such learning than we understand from interacting with real infants.

**NOTE:**
a sequence of short (e.g. 30 sec) video recordings (in mpeg format) of the antics of an infant born on 22nd Nov 2001 can be found here: http://www.jonathans.me.uk/josh/movies/ They provide a fascinating partial window into the mystery of an infant mind. Sound was added from about the end of October 2002. There are just over 30 movies at present (Nov 2002).

# A world of manipulables and communicating manipulators

- The robots will need to collaborate in solving problems
- They will be able to communicate with each other and with a human
- They will use a variety of speech acts including factual assertions, requests, questions, requests for explanations, explanations, warnings, etc.
- Gradually the complexity of the tasks will increase
- The robots will initially have to be programmed or instructed in order to extend their capabilities. Later learning and developmental processes will be added.
- Later on, various kinds of additional capabilities will be added to the system
  (perhaps eventually via a *general purpose* plug-in mechanism ??)
  - knowledge of various kinds of specialised subjects (such as those taught in schools)
  - various social skills
  - Teaching abilities (teaching reading, mathematics, programming, psychology...)
  - Entertainment and game playing abilities
  - Personal assistant abilities, etc. ...

# Affective control mechanisms

**One of the hard problems to be addressed is what motivational structures to build in to the robots: e.g., desires, goals, preferences, motive generators, conflict resolution strategies, etc.**

- For instance curiosity and a desire to help others will both be required, as well as desires to avoid harm and to end anything that is painful.
- Many sub-systems within the complete architecture may have their own desires, preferences, etc.
- Conflict resolution may be needed within or between any of the sub-systems.

# Why do it? Putting the pieces of AI together again.

## Like Humpty Dumpty AI fell off the wall. Can we find the king's horses or the king's men?

AI has become increasingly fragmented and factionalised since the late 1970s – partly because the number of people working in AI grew too fast. Many sub-groups started their own journals, conferences, mailing-lists, etc., and looked inwards.

It is not clear how much of the work done since then can contribute usefully to one of the most important goals of AI (and Cognitive Science) namely understanding how to design a complete, human-like (or even a chimp-like, or squirrel-like) agent, or equivalently(!), how to explain a large subset of the workings of a human mind.

That's because the vast majority of the work in the last 20 or so years has been focused on ever more narrow, though often deep and difficult, problem areas, without much thought about how to combine results from different sub-areas.

These slides sketch an ambitious project that aims to combine results from the sub-fields, in the hope that AI's fate can be better than Humpty Dumpty's.

## Can we work towards a complete, multi-faceted system?

# Social fragmentation vs content fragmentation

**This is not a proposal to undo the social fragmentation of AI.**

**I.e. There are far too many people working in AI now to form a cohesive social group. Moreover, many of the subgroups are doing excellent work which they can simply get on with. No one project can (or should) attempt to make them all interact in a rich way.**

**The aim is to undo the content fragmentation**

**I.e. the aim is to bring together a subset of AI researchers who want to think about designing "complete" and more human-like systems, especially systems which not only have various capabilities, e.g. being able to explain why they acted as they did, but also know they have some of them.**

**(At least we want to understand the problems, and find out what is and is not possible and why. What we mean by "complete" and by "human-like" will emerge gradually.)**

**Obviously we cannot design and implement a system like a human adult, or even like a human child.**

**So some abstraction and simplification will be required, as explained below.**

**Not everyone will be interested in doing this. There are many other worthwhile things to do! (Some of those will contribute to this project later.)**

# Disclaimer 1: There is no right way to do this

**Projects to re-assemble the components of AI could take many forms. No claim is made that the approach sketched here is the only one.**

**Different approaches could be pursued in parallel by different researchers,**

**AS LONG AS THEY TALK TO ONE ANOTHER. E.g.:**

- **Start with simulated robots that have a wide variety of competences, perhaps with a shallow implementation, and then gradually deepen the models and move towards a design that will work for physical robots**

  **OR**

- **Start with very simple physical robots and gradually try to extend the variety of physical and mental capabilities.**

  **OR**

- **Aim to produce something quite unlike a robot, e.g. a software simulation of a human legal advisor, or politician.**

**Our preference is for the first strategy, but we would be happy to collaborate with people adopting the second (or others with similar long term goals.)**

**Likewise we describe a particular sort of fairly rich spatial environment for the simulated robots, and particular sorts of tasks in those environments, but we would be happy to collaborate with people using different domains.**

**Different architectures and mechanisms could also be explored, in complementary projects moving towards systems that combine many human capabilities.**

# Disclaimer 2: Not everyone needs to do this

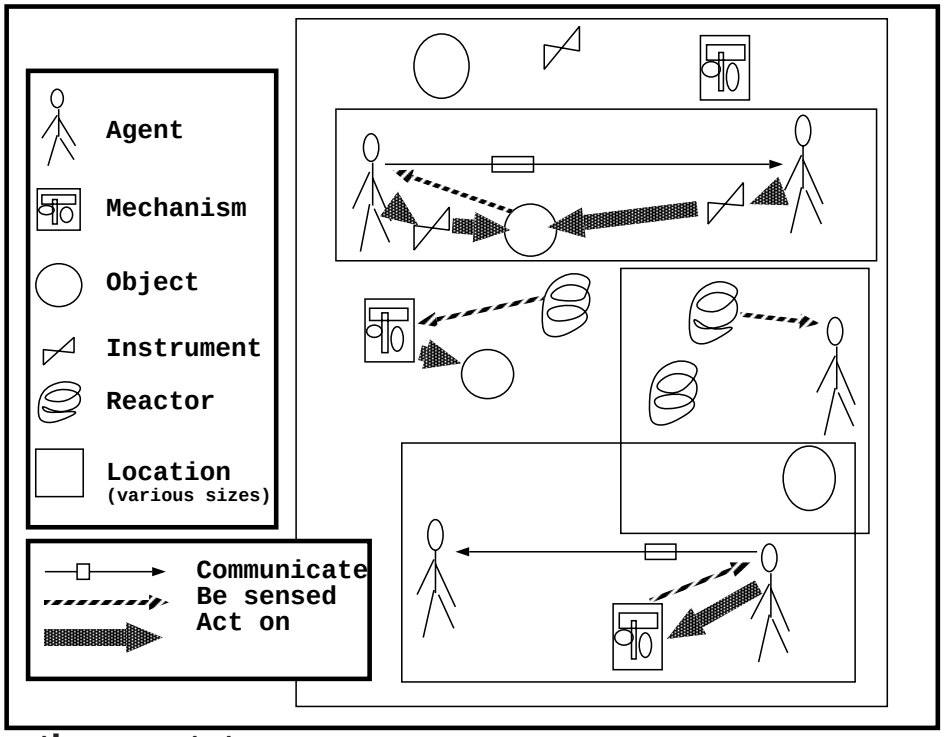
**There is no intention to claim that all AI researchers should drop what they are doing and join this sort of project.**

- On the contrary, progress in a project such as this will depend on continuing work on more narrowly focused problems of the sorts that have taken up most of the attention of AI researchers.
- Some of those projects may be influenced in useful ways if some of the people working on them become aware of the problems of fitting mechanisms, representations, processes into a larger, more complete, architecture.
- But it would be disastrous if everyone in AI immediately started trying to design complete systems, since the detailed technical progress could be slowed down.

**However, it will also be bad for the future of AI if too few people think about putting the pieces together.**

# What sort of environment?

## A high level requirement in the longer term: various kinds of concurrently active entities:




- AGENTS: which can perceive and act in the world, and possess some of these abilities:
  - communication with other agents (many kinds of communication)
  - planning
  - learning, development
  - forming of explanatory theories
- OBJECTS OF VARIOUS KINDS, E.G.:
  - MECHANISMS: which sense and react, e.g. a thermostat.
  - INSTRUMENTS: which can act if controlled by an agent,
  - REACTORS: which do nothing unless acted on (e.g. mouse-traps)
  - LOCATIONS: of arbitrary extents with various properties e.g. height, roughness, etc.
- Unobserved but possibly inferrable states and properties of objects, e.g. mass, friction, edibility, rigidity, strength
- Different kinds of causation
- Mental events, states, processes, mechanisms, etc. in agents.

# An example: the MJ world

- Two robots, M and J able to perceive, act, and communicate with each other and with a human (via terminal).
- M and J could have different physical and mental features, e.g. size strength, prior knowledge, goals, preferences. Possibly with different architectures.
  (Compare robots Mary and John, in Richard Powers' PhD thesis, Edinburgh circa 1973/4)
- A world in which there are several rooms and corridors, separated by impenetrable walls, except for doors linking some of them, which may be open or shut. Some may open only one way. Some may have keys that may be elsewhere. Some may be locked by sliding bolts. The bolts may be easily reachable or out of reach. Some walls may have windows. Later versions could have staircases.
- Various kinds of objects. Some large and immovable. Some small and movable. Some of them may be capable of being assembled to form larger objects: e.g. some blocks with holes, rods, wheels, capable of being assembled to form a toy car or trolley. (Compare assembly tasks set for Freddy, the Edinburgh robot in 1973.)
- M and J may have changing needs, e.g. for rest, food and drink. There will be suitable resting places and places with sources of food and drink. Curiosity, of many forms, will be a major motive.
- The world may be dynamic: e.g. there may be food sources that attempt to evade capture, and harmful agents that generate requirements to run, or hide, or fight. Physical objects may become unstable, and if large could be dangerous.

# What's a desirable and worthwhile goal?

**Specifying a level of HUMAN ability at any age is too ambitious.**
**Even a newborn infant is already learning much and has the capacity to grow into a bus conductor, or ballerina, or fish-monger, or a competent mother. We can't hope to emulate that.**

**So we shall have to abstract and simplify – as Galileo did.**

**We can try to identify a collection of capabilities of a young *idealised child of an imaginary species*.**

> **Even if many of the capabilities can be implemented separately fairly easily, designing them in such a way that many of them can be combined fruitfully, and adverse interactions can be controlled or used for learning, may be a very difficult challenge.**

**These capabilities will be of diverse kinds:**

- **sensing and perceiving**
- **desiring, needing, preferring**
- **acting and controlling**
- **inferring, deciding, learning**
- **wanting explanations and answers to questions**
- **communicating and explaining**
- **coping with interrupts**
- **developing over time**
- **representing and manipulating representations (a requirement for all the above)**

# McCarthy's Well Designed Child

**John McCarthy's paper "The Well Designed Child" lists important features of our environment, and some capabilities required to cope with them. See http://www-formal.stanford.edu/jmc/child1.html**

**He discusses kinds of innate knowledge and abilities, concerning:**

- the existence of persistent objects, forming natural kinds, with 3-D structure, locations, colours, relations, continuous motion...,
- kinds of situations, some of which recur,
- curiosity focused on information likely to be useful, treating other things as "probably noise",
- goal/sub-goal hierarchies and "the grammar of goal regression",
- the need for introspection, and the developing ability to do it,
- the "principle of mediocrity": I can learn about myself from observing others and vice versa,
- various kinds of meaning (to which grammar is secondary),
- the ability to process some kinds of information in parallel,
- the roles of logical and non-logical forms of representation (including use of chemical states to represent biological needs) – some of these express "virtual sentences",
- how to compose whole thoughts from components ("a word at a time"?),
- differences in linguistic requirements for perceiving, thinking (including use of pointers) and communicating, (Includes thinking about future possibilities.)
- varieties of reasoning using different forms of representations, including parallel inferences.

**How can we test McCarthy's proposals?**

# Disclaimer 3: We are not trying to design a complete child

## Some readers of an earlier version mistook the objective as being to build a complete model of a young child.

- It should be clear that that would be a *daft* objective: a human, at any stage of development, is far too complex to be modelled completely in the foreseeable future, perhaps ever, in an artefact designed by humans.
- What we are saying is that, as in all sciences, we can study some complex reality by abstracting from some of the detailed aspects in order to try to understand at least some of the more general features.

  (E.g. Think of Galileo and his balls rolling on planes, or Newton and his point masses).
- In the case of animals in general and humans in particular there many ways of doing this: e.g. it is possible to ignore mental phenomena and study animals only as physical systems, or only as units in an ecosystem.
- Our claim is that there is a level of abstraction at which we can usefully try to understand the interaction of many mental sub-systems that are not normally studied in combination. This is sufficiently demanding as to be at or beyond the edge of what can be done in the near future, and yet not as wildly ambitious as trying to model a complete body, including brain from the atomic level upwards, for example. (But our objectives are still too vague.)

# Disclaimer 4: We don't disparage more "bottom-up' research

- Some people may wish to model muscles, neurons, metabolic processes.
- Some may start from neurons and chemical circuits in brains, and try to find out what can be built on top of those.
- Some may start from existing AI achievements (various kinds of learning systems, analogical reasoning systems, planning systems, vision systems, language systems, etc.) and try combining them to achieve specific objectives, or to find out which combinations can do something useful.
- These alternative approaches may provide useful results.
- But attempting to understand various kinds of missing functionality in complete AI systems and work backwards towards required architectures, mechanisms, representations, etc. may also be useful.

**Not everyone needs to work in the same way. All we are trying to do is find potential collaborators – and constructive critics.**

# The innate/developed/learnt trade-off

**Given a list of kinds of knowledge and abilities for a "normal" mature individual (human or robot) it will not be clear how much needs to be innate, and we don't yet know all the trade-offs.**

**Typical mature individuals could be implemented using decision nets listing appropriate responses to all possible sequences of environmental events. But this would be impossible in practice because the net would be too big to be built in our universe and because no designer could anticipate all possible futures.**

**One of evolution's answers:**

- **Some species need enormous flexibility and creativity. They use powerful innate bootstrapping and learning mechanisms ("altricial" species – including the most intelligent mammals and birds.)**

  **This is heavily dependent on the availability of a suitable environment in infancy, and requires detailed parental intervention, AT LEAST FEEDING AND PROTECTION, AND IN MORE SOPHISTICATED CASES EDUCATION AND TRAINING.**

- **Other species have less demanding requirements: their behavioural capabilities are all (or almost all) selected by evolution and encoded in genes ("precocial species" e.g. insects, fishes, ...)**
- **Various intermediate forms exist, e.g. many less intelligent mammals, birds, reptiles, etc.**

**We do not yet know all the design options, and the trade-offs between what needs to be innate and what can be learnt. So our project may have to explore different options to find out how well they work.**

**Compare the Kantian arguments for innate concepts in Talk 14 here:**

**http://www.cs.bham.ac.uk/~axs/misc/talks/ (attack on symbol-grounding)**

# A possible robot

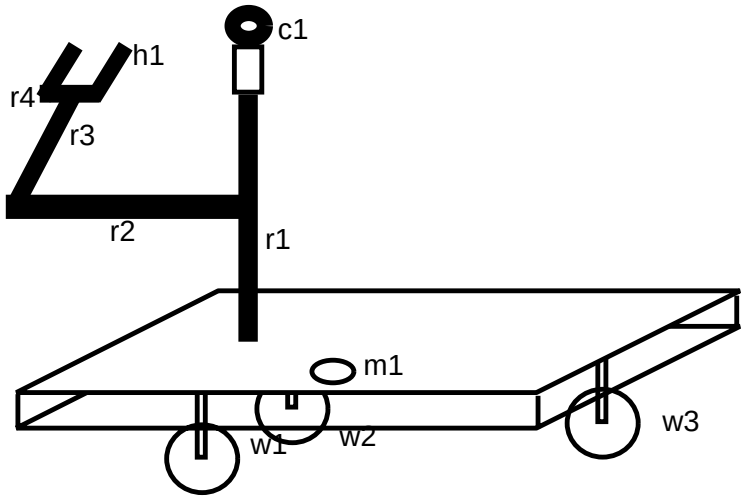
**Individual robots can vary in complexity and in "realism" e.g. slippage, inertia, etc.**

**Initially the robots may have a very simple structure, very few degrees of freedom, relatively few modes of manipulation (e.g. only the ability to push objects from one place to another.)**

**Later they can be articulated, with more complex capabilities and needs.**

**The picture shows one of medium complexity.**

**Another sort of option would be something like the Sony AIBO dog with a mouth that can grip things.**



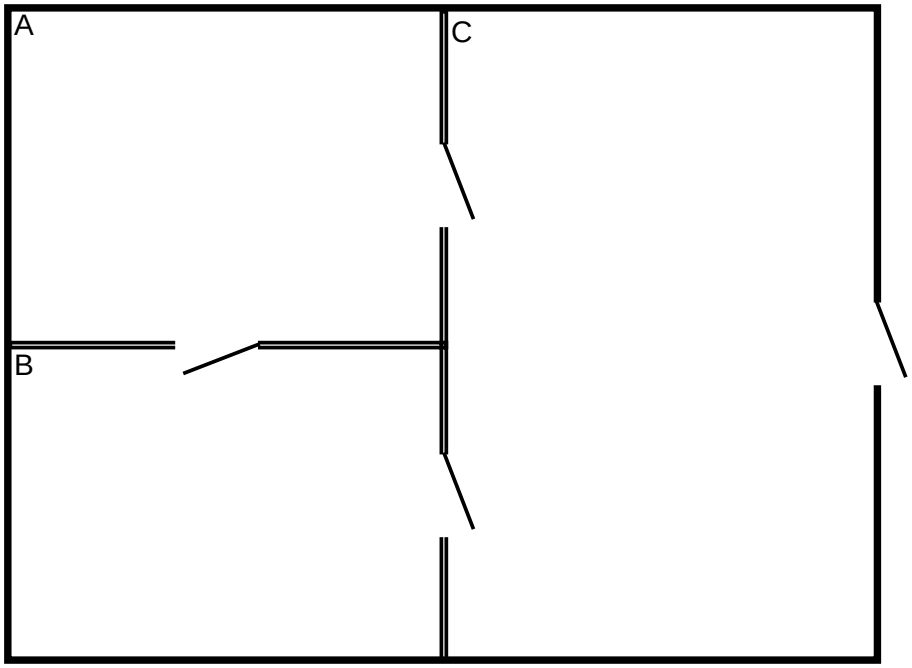

r1 can rotate, causing r2, r3 h1 to rotate in horizontal plane
r2 can move up and down r1, causing r3 and h1 to move vertically
r2 can rotate causing r3 and h1 to rotate in vertical planes
r2 can shrink or stretch, causing r3 and h1 to move horizontally
(or should that be left to wheel movements?)
r3 can rotate, causing the plane of r4 and h1 to change
r4 can rotate causing the plane of h1 to change
h1 can move its fingers closer together or further apart
c1 is a camera: can turn, tilt and adjust focus (independent of r1?)
w1 w2 w3 can rotate, causing motion. w1 and w2 rotate at the same speed or can free wheel, while w3 can be driven independently.
w3 can also rotate about vertical axis.
m1 (mouth) opening for food and drink to be poured in?

Proprioceptive and force sensors: r1 r2 r3 r4 h1 and wheels.
Also weight on platform
(Weight increases difficulty of moving. If too heavy cannot move?)
Internal state sensors: hunger, thirst, need to rest.

**Many existing robots are incapable of perceiving similar robots in action, e.g. seeing themselves in a mirror, unless they are very simple robots.**

# A possible environment for two robots

The project could have some tasks involving manipulating objects in full view of the robots.
Other tasks may require moving round a larger environment, not all parts of which can be visible at the same time.
These movements may or may not be hampered by locked doors, keys to be fetched, a need for co-operative action to open a door (e.g. key on one side of door, but hole for insertion on the other side, or one robot providing a platform for the other to reach a key, etc.).




**As the project grows we can alter the scale of the terrain, the variety of features in the terrain, the complexity of the tasks required to move from one part to another, etc.**

One representational task is linking spatial information on different scales: understanding how what is seen is part of a larger space going beyond what is seen, i,e, merging perceptual data-structures with long-term knowledge structures.

This uses the idea of zooming in and out when thinking about a spatial region.

# Spatial competence - 1

## At an intermediate stage, not necessarily the very first stage, the robots should have something like the following set of capabilities:

Robots can move forward, backward, turn left or right, rotate on the spot.

Arms can move up, down, change geometry to alter reach.

A hand or gripper can open and close and change its orientation, perhaps in two dimensions.

Robots can pick things up, push things, place things on platform for carrying, etc.

Camera has motions independent of hand and arm and can focus on the same thing at different distances.

Perceptual mechanisms could produce both raster arrays (or perhaps 'polar' arrays: concentric rings of variable sized receptive fields) and structured descriptions – something like parse trees. Initially we need not care how this is done. Later we'll have to address functions of vision more carefully, including the perception of affordances.

## Early versions will have simplified capabilities.

E.g. to start with, some actions could be achieved in the simulated world by 'magic' without fine control of motion of components of hand-arm system. Later finer grained simulation must be added. Likewise perceptual magic, at first.

# Spatial competence - 2

## The spatial competence should allow various kinds of integration

- Integrating different views of the same scene following eye movements (camera-movements) of various kinds. while the rest of the body does not move.
- Integrating different views produced by movements of the body (i.e. changing viewpoint as well as changing view direction).
- Integrating different views of the same scene over time, e.g. as perceived objects move.
- Integrating information from different sensory modalities (e.g. touch sensors) and motor output.
- A later version could include binaural auditory input providing spatial information.
- Visual and other spatial perception will not merely produce information about physical properties (spatial structure, relations, orientations, textures, colours, etc.) but will also produce information about affordances (i.e. possibilities for action relevant to the agents goals, preferences, capabilities, location, and constraints).

# Warning: start simple & use stepwise refinement

**Evolution produced very impressive multi-functional complex organisms, by going through stages in which simpler but complete organisms developed, whose capabilities were re-used in more complex systems.**

**This happened at many levels including re-use of mechanisms for growing and maintaining individual cells, mechanisms for maintaining temperature, mechanisms for producing directed movement, mechanisms for controlling posture, etc.**

**This involved *a succession* of fitness requirements.**

**If we try to design something too complex all at once, we may fail to identify an appropriate collection of layers of competence.**

**So this project should at least mimic evolution in developing a succession of complete, functional systems, adding layers of functionality which build on previous kinds of competence where possible (sometimes using the 'duplicate then differentiate' strategy often used in evolution). However it is impractical to try to mirror details of animal evolution, apart from high level stepwise refinement.**

**Conjecture: a collection of abilities to perceive, reason about and manipulate spatial objects, processes and their relationships provides a powerful infrastructure for many other kinds of abilities.**

# One kind of refinement: growth of fluency

**Often people will start performing a certain multi-step action (e.g. starting a car and driving off, or playing a musical instrument) in a fairly slow and jerky manner, doing one step at a time.**

**Later they develop a fast and fluid performance.**

- **This may be done by a deliberative system somehow training a reactive system to perform the steps quickly, via a succession of trained reactive associations.**
- **If this requires some lower level continuously varying parameters, then their variation could be modulated by another part of the architecture, to produce a controlled performance.**
- **A different sort of modulation would produce a fast and smooth performance.**

**All that could be achieved by a robot which had little or no knowledge of what it was doing or why.**

**Combining that with the sort of self-knowledge required for learning to anticipate problems, or for communicating methods to others might enable a robot to explain its actions.**

# Isn't building simulated robots cheating?

**ANSWER:**

**anything doable in the near future will involve cheating (simplifying)**

**Different types of "cheating":**

- **Building physical robots that move using wheels rather than legs, fins or wings.**
- **Building physical robots that lack the ability of humans and chimps to grasp and manipulate objects with fingers**
- **Building physical robots that cannot make use of high resolution perceptual input in complex cluttered scenes with structured objects.**
- **Building physical robots that lack the detailed physiological structures and metabolic processes of animal bodies**
- **Building physical robots that use computers instead of biological neurons**
- **Building simulated robots**

**CHOOSE YOUR PREFERRED TYPE OF CHEATING!**

**Are designers of air-liners cheating when they test important design features using simulations?**

**Not if they know what they are doing – and its limitations.**

**Compare the simulations used in the RoboCup and RoboCupRescue competitions.**

**http://www.robocup.org**

**http://robomec.cs.kobe-u.ac.jp/robocup-rescue/**

# Conjecture

**Suppose we can give the robots a fairly rich and deep understanding of space, time, motion and causation (including an understanding of actions the robots can perform), along with understanding of goals, plans, successes, failures, explanations, collaboration, helping, hindering, etc. in the simulated world.**

**Then: that will provide a basis for developing a wide range of problem-solving and communicative (conversational) abilities.**

**What sort of understanding of the world is required?**

**That is to be answered by defining and implementing systems capable of demonstrating an increasingly complex and varied collection of abilities, in scenarios of types indicated (very roughly) in these notes.**

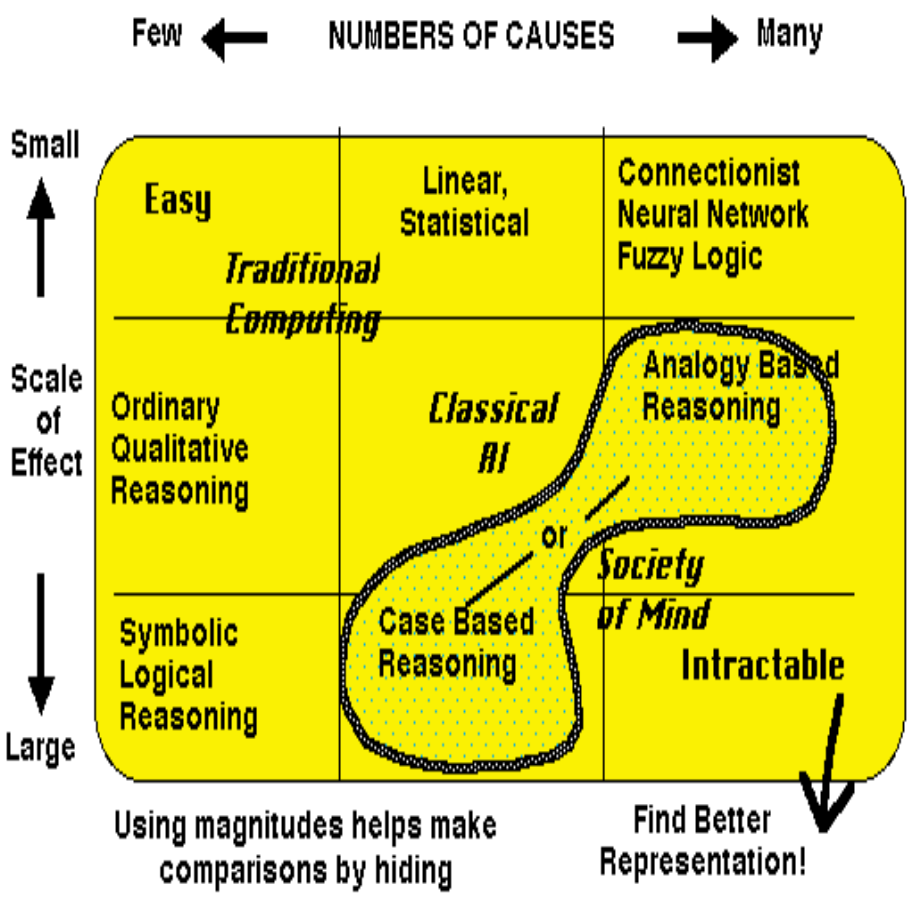
**(To be clarified and justified.)**

# Minsky on varieties of representation

**There has been much debate on whether logic suffices for AI.**

**Many have argued that different forms of representation are needed for different tasks.**

- **logic (specially good for specification)**
- **diagrams/spatial forms**
- **graphs, trees, etc.**
- **rule-sets**
- **procedures**
- **neural nets (various kinds)**
- **states in a dynamical system, etc,**




**In his paper 'Future of AI Technology' Minsky represents some of the design options for representations to be used in a reasoning or problem solving system, as shown in his diagram. See http://www.media.mit.edu/people/minsky/papers/CausalDiversity.html,**

**We need more analyses of trade-offs between different architectures, different forms of representation, different algorithms, different physical designs.**

**We can expect our robots to use a variety of forms of representation.**

**My papers on varieties of representation are here: http://www.cs.bham.ac.uk/research/cogaff/**

# Do organisms use representations, or compute?

## Two confused debates:

- Sometimes people use a very narrow notion of 'representation', and argue that organisms (including humans) do not use representations, at least not in their brains, and therefore robots do not need them either.
- Sometimes people use a very narrow notion of 'computation' e.g. based on bit-patterns or Turing machines, and argue that brains therefore do not perform computations, and perhaps therefore robots need not.

**Both arguments depend on the assumption that the notions of "representation" and of "computation" are well defined and well understood.**

**In fact both are commonly used in confuse and confusing ways.**

**E.g. see**

http://www.cs.bham.ac.uk/research/cogaff/Aaron.Sloman_towards.th.rep.pdf
http://www.cs.bham.ac.uk/research/cogaff/sloman.turing.irrelevant.pdf

# Programming language requirements

**Simple versions of the robots may need only a mixture of condition-action rules, along with standard AI language facilities, e.g. list processing, math functions, arrays, pattern-matching etc.**

**The flexibility, rapid-prototyping and interactive development capabilities of AI languages may prove more important than features of more widely used languages, e.g. efficiency or portability.**

**The rule interpreter should (a) allow conditions and actions to invoke arbitrary language functions, (b) be multi-threaded, making it easy to implement**

- **hybrid interacting mechanisms (e.g. neural and symbolic systems)**
- **concurrently active agents and objects**
- **concurrently active mental components within individual agents**
- **perhaps the ability easily to vary the relative speeds of different threads**

**The sorts of implementation mechanisms proposed in Minsky's *The emotion machine* could probably all be implemented in such a rule-based system. However it may be desirable to have special-purpose implementations for efficiency.**

**Other special purpose mechanisms may be required later, including planners, visual mechanisms, data-bases, language mechanisms, physical simulations.**

**All will be justified by the scenarios they support well, not by religious warfare about languages.**

# A robot architecture

**One research strand could use variants of the Minsky/Sloman architecture scheme, elaborated as needed over time – e.g. see the H-CogAff slide below.**

**An intermediate implementation might include:**

- Some innate reflexes, e.g. to prevent damage
- Some innate reactive behaviours which include internal state changes (e.g. Nilsson's teleoreactive systems).
- Some adaptive reactive mechanisms (including trainable "alarm" systems).
- Some simple deliberative capabilities
- Some meta-management/reflective capabilities (needed for learning, explaining).
- Various kinds of goal generators, preferences, and other affective states and mechanisms, including arbitration mechanisms.
- All the above operating concurrently, with interrupts, etc.
- Simple hierarchies of perceptual and motor capabilities.
- Simple linguistic capabilities
- A store of innate information about the physical environment (e.g. an ontology), along with mechanisms for extending it. (Cf. McCarthy's papers.)
- A store of innate information needed by the meta-management system including an ontology for internal processes of various kinds.
- Later: some abilities to learn or develop the above.

# Thought vs Action

Much AI work is about knowledge, representation, inference and learning, addressing the task of discovering and representing what is the case, including the ability to answer questions.

In animals and robots there is also a requirement to *do* things. This includes both external and internal behaviour. What are the springs of such actions?

- In the simplest cases there are no goals: merely reactions. A car is designed to react to a combination of externally produces changes to steering wheel, brake pedals, accelerator, etc.
- Slightly more complex machines accept a single top level goal, from which sub-goals, and ultimately actions, are derived.
- For animals and autonomous robots, goals mainly come "from inside", though they may be triggered by external events (e.g. a perceived danger or opportunity to eat).

Animals and sophisticated robots have externally triggered goals in a dynamic environment along with a body that has changing needs (fuel, lubrication, repair, protection from damage, relief of stress, temperature control, etc.).

They require mechanisms that can handle multiple, dynamically changing, goals with varying priorities and importance, in various combinations.

Initially we try to minimise the need to cope with this problem by simplifying the design of the robots and the environment.

Later, we shall have to produce an architecture that copes with many types of motivational and affective processes.

# Affective control mechanisms

- **Motivation**
  Basic requirements for survival, plus desires to help, curiosity, a desire to understand the environment (including learning generalisations and acquiring deep explanations); a desire to solve problems or perform tasks efficiently and elegantly; to impress or outperform other agents)
- **Emotions and moods**
  Instead of being based on specific "emotion" mechanisms, emotions on various time scales may mostly be emergent states arising out of interactions between motivational and other control mechanisms, including more or less global "alarm" mechanisms. Mood control may not be needed in early prototypes. In more advanced versions environmental changes may induce moods.
- **Attitudes**
  Mostly dormant collections of learnt motivations, preferences, beliefs and strategies, triggered by specific events. Not needed in early prototypes.
- **Arbitration mechanisms and rules**
  Preferences, priorities, and a variety of arbitration mechanisms will be needed for dealing with conflicts of motivation and decisions, and for combining complex concurrent actions with inconsistent preconditions or tendencies.

**All this is just scratching the surface.**

# Beware of published theories of emotions

**Many people think that producing intelligent systems will require including emotions somehow. Too often they think of emotions as "skin deep" (visible behaviours and physiological state changes).**

- **Most emotion theorists don't know how to explain states in terms of what's going on in an information-processing architecture.**
- **So they try to characterise emotions in terms of observable/measurable phenomena (blood pressure, galvanic skin response, weeping, smiling, ... etc.)**
- **Instead we should develop an affective ontology based on varieties of control processes within an architecture, e.g. generation of motives, detection and resolution of conflicts of motives, priority changes, alarm mechanisms and interrupts of various sorts, switches of attention and modes of processing, etc.**
- **In various papers I've shown how at least three major categories of emotions are associated with the three types of architectural layers (primary, secondary, tertiary emotions), but that's a grossly over-simplified taxonomy.**

**A full theory would start from our common-sense taxonomy of desires, preferences, pleasures, pains, values, ideals, attitudes, concerns, interests, moods, emotions, intentions, etc. and then produce a richer, more precise, architecture-based taxonomy of affective control states.**

**Compare: H.A. Simon 'Motivational and emotional controls of cognition', 1967 (repr 1979 in *Models of Thought*, Yale Univ. Press)**

# The need for scenarios

**In order to drive the design process we need hypothetical scenarios, i.e. "film-script" descriptions of demonstrations that could be given after the project has made significant progress.**

- Initially we'll be able to produce only fragments of scenarios.
- Gradually we can tie them together into larger "film-scripts".
- We may need to produce several hundred scenario fragments to help focus our ideas.
- Then we can work backwards towards architectural and representational requirements to make those scenarios work, as well as identifying the required knowledge.
- The aim is not to hard-code each scenario separately, but to identify a collection of them whose requirements overlap sufficiently that by ensuring that they are all supported in an elegant and economical fashion we ensure that we have a very general set of capabilities that will also support a wide range of additional scenarios.

  Compare data-mining over a structured database, or inducing a grammar from a set of sample sentences.

- For many scenarios we should also develop meta-scenarios in which the agents talk about, answer questions about, or explain what they have done or could have done, instead of merely being able to do it (like many non-human animals).

# A scenario fragment: how to grasp

**Robot A is trying to get robot B to assemble a toy car (as Freddy did, in Edinburgh, 30 years ago):**

**A Push rod 5 through hole.**

**B Why?**

**A Because: rod makes axle.**
**AND wheels will go on ends of rod**

**B OK.** **(What are the requirements for A to give that explanation and for B to understand it?)**
**...grabs middle of rod and picks it up...**

**A Stop.**

**B Why?**

**A (Put rod down on table) then (grasp rod at end), not (grasp at middle).**

**B Why?**

**A (Grasp at middle) hinders (push rod through hole).**

**B OK. ...puts rod down...**

**B Which end of rod to grasp ?**

**A Grasp end nearest hand.**

**B OK**

**... Later B finds he can insert the rod in two movements if he grasps it in the middle to push it into the hole. What are all the things he learns from that? E.g. problems have alternative solutions, with trade-offs?**

# Scenario involving a bargain

**Consider this conversation:**

**A: My fuel supply is low. Could you go to the fuel station and get me some?**

**B: I want to finishing building this bridge.**

**A: I can't survive till you have finished.**

**B: Why don't you go?**

**A: I don't have enough fuel to get there.**

**A: I'll get more bricks for your bridge while you are going for fuel.**

**B: OK I'll go.**

**Towards an analysis:**

**A has a goal but cannot achieve it, so works out a plan involving B. (How?)**
**Acting on the plan is inconsistent with one of B's goals. (How does B detect the inconsistency? How does A understand it?)**

**B produces an alternative plan not knowing that it is not feasible.**

**A interprets this as indicating that B will not help unless rewarded. (How does A draw this inference? Could it be a result of learning – about modes of argument?)**

**A makes a plan that includes A rewarding B., etc. (What enables A to come to think of that plan? Why does A choose that rather than some other?)**

**What is it about B's motivational structure that leads B to agree to go?**

# Analysis of scenarios

**We can analyse the significance of such scenarios by identifying the kinds of knowledge and abilities that they demonstrate and the kinds of problems our robots (and our children) need to be able to solve.**

**Examples (in no particular order)**

- **Instructions/requests can be obeyed blindly or with understanding of reasons. Reasons can be requested and provided (sometimes)**
- **Any particular goal can generally be achieved in multiple ways, and different means will have different consequences, often involving unobvious trade-offs.**
- **One agent's prior learning regarding such trade-offs can be used to optimise another agent's actions.**
- **There's also a large collection of issues to do with how perception of physical situations provides information about possible actions and constraints on possible actions (Gibson's "affordances").**
- **The scenarios as presented could occur in a variety of physically (geometrically) different configurations of objects which all essentially present the same high level affordances. Giving the robots an adequate understanding of that generality may be very much harder than enabling them to handle a collection of special cases.**

# Geometrical structure and causal powers

**In some scenarios we may need to have our robots understand things like the following:**

## Types of long thin things:

- Consider thin rods of the same thickness all made of the same material: as they get longer the rods will be harder to move without breaking them.
- The material a long thin object is made of will determine whether it is (a) bendable, (b) stretchable, (c) compressable, (d) easy to break, (e) able to return to its shape after being bent or stretched, (f) easy to sharpen at the end
- Some long thin things can be used to tie up parcels. Others can be used to push things out of long tubes. Others can be used to "fish" something out of a hole by spearing it.
- some long thin things are graspable and movable, others not (e.g. a long thin shadow, or stream of water coming out of a pipe).
- the longer or the thinner an object is the smaller the weight of an object it can be used to push without bending or breaking.

## Types of visual obstruction:

- The size of an object will determine how much of another object on the far side it hides.
- The shape of an object will determine which parts of another object it hides.
- As you move closer to an object the more it will obscure of other objects on the far side
- Moving sideways or up and down will alter which portions of objects on the far side are obscured.
- if an object moves its location or changes its shape that can alter how much or which portions of other objects it obscures.

## How does a child understand these things? What is this "understanding"?

# Variants of the same competence in an individual

**Within an individual the kinds of knowledge and capabilities described previously may have different forms, used in different contexts.**

- **Skilled action may involve trained responses within a reactive subsystem which uses perceived patterns to generate appropriately nuanced physical actions (pushing, pulling, grasping, bending, etc.) in different contexts. The individual need have no explicit knowledge of what's going on, and may even be able to perform these actions in parallel with deliberative processes.**
- **For some complex actions, explicitly formulated symbolic plans may be used, including instruction sequences communicated by others. This can produce action sequences that work, but are not necessarily fluent and, which use cognitive resources not required in the former case of fluent reactions.**
- **The ability to understand why actions work, predict when they will fail, answer "why?" questions, etc. may require additional meta-management capabilities and higher order representations, which do not develop till later. (How?)**

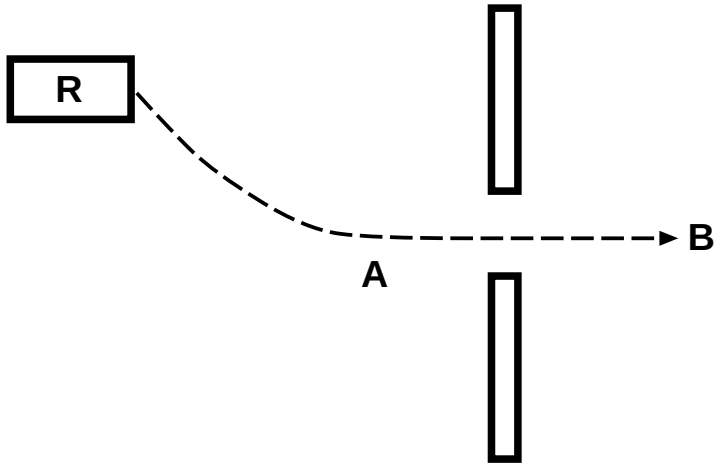
**Competent robots (whether physical or simulated) can include different subsets of the above capabilities: our goals need to be clear. In humans the three sorts of competence probably become available at different ages. The third, reflective sort may be closely related not only to communicative abilities but also to the development of mathematical abilities, using new powers of abstraction.**

# Meta-scenario: going through a door

**A rectangular robot R can move quickly and smoothly through a rectangular doorway D which provides little clearance, to a target point B. It might be good be good to make the move in two stages, first to point A, then on to B.**

**Getting a robot do all that is one thing.**

**Getting it to understand what it is doing, and why, requires more.**




**M: Why did you head for A before turning: why not join the line AB at a closer point?**

**J: Because that would have made the journey longer.**

**M: Then why did you not make the journey even shorter by heading for a point even nearer the door before turning?**

**J: Because it would have been more difficult to change orientation in time to get through the door without hitting anything.**

**Are there robots that have that kind of understanding of their actions, which they can then use in describing, explaining, or defending their actions, or in understanding the actions of another, or in teaching another?**
**(Or teaching a human?)**

**At what age might a child understand this? Could that sort of understanding play a role in understanding a story about a collision in a doorway?**

# Varieties of geometric construction kits

**Construction kits for children come in diverse forms. Many children learn to "play" with different sorts of kits, with different kinds of geometrical and causal properties.**

- **Mouldable stuff with continuously changeable shape, but no rigidity or tensile strength, e.g. making sand castles and mud pies.**
- **As above but with some tensile strength and limited rigidity, e.g. plasticine, putty, chewing gum. (What about liquids?)**
- **Discrete stackable objects with flat surfaces: blocks**
- **Rectangular "locking" shapes, e.g. LEGO bricks.**
- **Objects with holes and rods that can go into them making rigid or loose connections, e.g. tinker toys.**
- **Objects that can be combined so as to make non-rigid joints between rigid components, with screws or clips holding things together, e.g. meccano.**

**Has anyone tried to unravel all the knowledge, the perceptual skills, the motor skills, required to assemble toys with all of these sorts of kits? Compare being able to give explanations how why things work or don't work.**

**What sorts of mechanisms are required to learn these things? Could it all be programmed explicitly, or must some significant subset be learnt? Why?**

**Is there any other animal that can learn to handle all these? If not why not?**

# Construction scenarios: different requirements

**The investigation of different sorts of "construction kits", whether naturally occurring (sand on beach, mud) or artificially designed (Lego, Meccano) may enable us to understand the different cognitive/developmental requirements for using them. For instance:**

- **They vary according to whether assembly operations are discrete (add another brick) or continuous (pour on more sand) or a mixture.**
- **They vary in the minimal number of components required for an assembly task: two for lego (a brick can be pressed onto something) and four for meccano (two objects with holes, nut and screw).**
- **They vary in the extent to which modification is possible after addition: rigid components prevent modification, bendable components may be accommodated after combination.**
- **They vary in the kinds of functions constructed objects support, e.g. merely having a certain shape versus allowing certain classes of motions - wheeled vehicles, cranes, etc.**

**Some hard questions: How much of the ability to cope with all the varieties is innate? How much of it develops out of prior construction activities? In that case what changes: is it just addition of knowledge, subroutines, or is the architecture changing through action? Can we model this?**

# Non-physical construction kits

**Being able to construct toy houses, dolls, bridges, cranes, animals uses the ability to manipulate physical objects in an intelligent way.**

- **In animals, that uses enormously intricate biological mechanisms including complex chemical and physiological structures and processes. It may be that some aspects of the control problems are independent of that kind of detail. We need to find out – not strike dogmatic attitudes.**

**Being able to construct complex percepts, thoughts, motives, plans, memories, questions, explanations, etc., need not involve any external communication using words, pictures, gestures, etc.**

**The internal processes require some sort of mental construction kit, with components that can be stored, re-used, combined in different ways, and structures that can be explored, evaluated, rejected, etc.**

- **Animal brains use enormously intricate mechanisms including complex chemical and electrical processes, and perhaps processes of types we do not understand.**
- **Perhaps some aspects of those abilities can be replicated using a different low level physical infrastructure, and some cannot. Which can? Those that are essentially virtual machines? We need to find out – not strike dogmatic attitudes.**
- **It may be that the main features of the architecture can be replicated electronically even if many fine-grained details cannot.**

# Learning to use a mental construction kit

**Gerald Sussman worked around 1971 on a program called HACKER which could construct plans for achieving simple tasks by moving blocks around.** (MIT Phd Thesis, later published as a book.)

- Some of its plans were buggy. When bugs turned up, the program had ways of inspecting the execution process in order to determine the mismatch between expected results of plan steps and what actually happened.
- This led to discovery of high level patterns in the interactions of plan steps. (Example, actions to achieve one sub-goal undo results of actions to achieve another sub-goal.)
- Some of these newly discovered patterns of buggy plan execution were reflected in patterns that existed in the process of plan construction: syntactic clues to semantic properties.
- So HACKER could modify itself by creating new "critics", monitor programs which could watch the process of plan construction and interrupt if signs of a bug occurred.

**When does a child start de-bugging during plan construction as opposed to de-bugging during plan execution? Can any other animals do this? How important is this in a teacher? Which human problems arise out of failure to learn to do this (e.g. in social plans)?**

# Conversational competence and "physical" competence

## Conjectures:

- Insofar as children (or our planned robots) have the ability to talk about what they are doing, including not only describing it but explaining why they perform particular actions, and explaining why doing it differently will not work, they will need meta-management (reflective) mechanisms in their architecture.
- Having this ability is also connected with the ability to observe the actions of another, and give help and advice to another.

  (I.e. some social skills are architecture-based).

  Qualification: a "dumb helper" may be trainable so as to give useful advice in Eliza-like fashion without understanding the questions or the answers.

  But we are not aiming at that kind of tool.

  Why not?

# Construction scenarios: Adding or removing a dimension

**A young child can be given a meccano set, and a 2-D picture of a 3-D object made out of meccano pieces.**

**We could have scenarios where something is presented as a 3-D construction and the task is to make another like it, and scenarios where only a picture of the end result is given, and the task is to build something depicted by the picture.**

**Different variants of the scenario:**

- **one robot solving the problem alone**
- **two robots solving the problem in collaboration**
- **a learner robot being helped by a person or a teacher robot who does not manipulate the components, but has to give the learner suggestions, answers to questions, explanations.**

**What capabilities does the second robot (or person) need in order to be able to judge when it would be useful to intervene, either by asking a question or making a suggestion, or just making a comment, e.g. drawing attention to some fact: "Later on what you have constructed will over-balance".**

**I.e. what features of a process of action can be seen as an indication that the problem has not been fully understood? Which features indicate what kind of intervention will be most helpful for the learner?**

**These scenarios include continually changing affordances, both in the unfinished construction and in the picture guiding the construction. How does a teacher know which ones a pupil has grasped?**

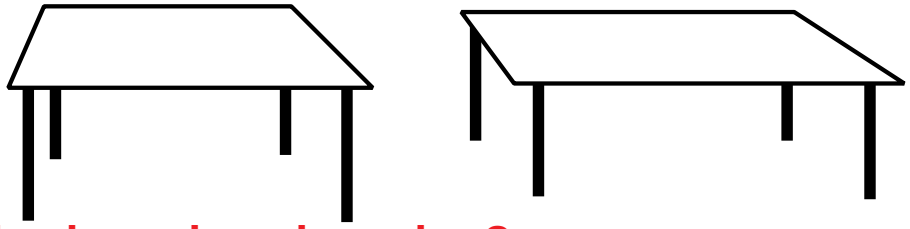
# Construction scenarios: some issues

- The previously discussed questions about what should be innate, what should be learnt, and what should come from a genetically determined process of development guided by the environment will frequently recur.
- Pat Hayes' 'Naive physics manifesto' is very relevant
  e.g. in *Formal Theories of the Commonsense World*, eds. J.R. Hobbs & R.C. Moore.
  See also http://ontology.buffalo.edu/smith//articles/naivephysics.html,
  which includes some history of ideas about "naive physics" before AI.
- Where different sorts of physical components, or different sorts of construction kits are available, how does this affect the requirements: forms of perception, forms of representation, forms of motor control, etc.?
- Where robots or animals with different kinds of physical manipulators do the construction, how do the differences in their bodies affect the differences in their mental requirements?
- To what extent is the ability to *describe* actions and *explain* them based on the abilities involved in *performing* the constructions, and to what extent do they require separate forms of representation and reasoning?

# Some issues regarding vision

**Vision is perhaps the hardest problem in AI, on which we still understand very little. It has multiple facets.**

**How do we get from 2-D patterns of illumination on our retinas to percepts of a 3-D world:**

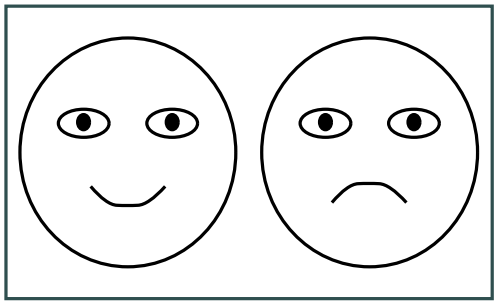


**What if the objects are flexible, irregularly shaped, and moving?**

**How do we see expressions of emotion in faces?**

**How are emotions represented in perceivers?**

**Which begs the question: What are emotions?**

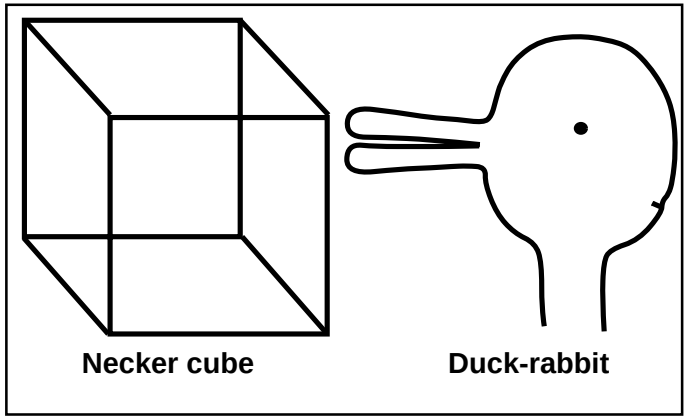


**How can we see the same 2-D visual input in different ways?**

**What's the difference between a geometric ambiguity and the duck-rabbit ambiguity?**

**How are the differences, e.g. between "looking left" and "looking right" represented in perceivers of perceivers?**

**And many more, including perception of motion, perception of causation, of affordances, of functional relationships, of actions, etc.**



# Analogical reasoning scenarios

**A once very famous program solved geometric IQ test problems of the form: A is to B as C is to which of pictures D,E,F,G?**

**T.G. Evans, 1968, A heuristic program to solve geometric analogy programs, in *Semantic Information Processing*, Ed. M.L. Minsky, MIT Press, Cambridge, Mass, pp. 271–353.**

**This required the ability to describe structures, to describe relations between structures, and to describe and compare relations between relations between structures: e.g. which of the relationships C to D, C to E, C to F, C to G, is most like the relation A to B? The answer to that can also depend on comparing the relations between A and C, with relations between B and D, E, F, G.**

**X: A is to B as C is to F.**
**Y: Why?**
**X: Because if you take the triangle in A as corresponding to the triangle in B, then you can take the square in C as corresponding to the square in F.**
**Y: But what about that arrow in A and B?**
**X: It is not relevant as there is no arrow in C.**
**Y: But if you notice that the arrow in A points away from the middle of the triangle in A and the arrow in B points away from the middle of the triangle in B, then we can see the two pointed shapes in C and E as playing the same role in relation to the squares in C and D as the arrows in A and B play in relation to the triangles in A and B. So I say the answer is E, not F. etc.**

# Analogical scenarios

**Suggestion:**

**We should look at a lot of cases of analogical reasoning (AR) in order to find out how many kinds there are, which different capabilities they use, and what role they play in perception, problem solving, learning, explaining. (We could try using existing AR algorithms.)**

- **AR is probably crucial for meta-management: i.e. being able to monitor, evaluate and improve one's own thought processes – for that involves seeing relationships between relationships,**
  - **e.g. action A1 produced effect E1 so what effect will action A2 produce?**
- **Compare and contrast the kinds of analogical reasoning that use structural descriptions with those that use neural nets.**
- **Is there a good overview of relevant literature that could be applied to our two robots in the manipulable world?**
- **Compare investigations of how metaphors work in linguistic communication. (The ATT-Meta system of Barnden et. al.)**

**Analogies, metaphors, using one thing as another (a knife as a screw-driver) and counter-factual conditionals all involve changing some things in a complex situation but not others. We need to understand the many ways this can be thought about, and how the mechanisms for doing such things develop in a child.**

# Mental control and attention

**In a complex and changing world, one of the tasks a child has is selecting what to attend to. (Why can't *everything* be attended to at once?)**

**The variety of failures and disorders of attention found in humans indicates the complexity of attentional mechanisms. Will similar complexity be required in our robots?**

**We regard *attending* as *selecting among alternatives*.**

- **Selecting a direction in which to gaze, or turn one's ears.**
- **Selecting which feature of an object to look at (colour, texture, structure....)**
- **Selecting which consequences to think about**
- **Selecting between perceptual information in the external environment and reflection on internal processes.**
- **Selecting between different high level tasks (planning an essay, remembering a pleasant experience, thinking about where to go for dinner, wondering whether a train will be late, contemplating a possible mate, finding food ...)**
- **Selecting between methods for performing a task, e.g. trying to remember an answer, trying to work out an answer, looking up the answer in a book, etc.**

**How much of this applies to insects, to dogs, to chimps, to new-born infants, to a three year old child? What kinds of attention control will our robots need?**

**Different forms of attention control are needed in different architectural layers.**

**Two aspects of attention control: selecting, and maintaining the selection.**

# Problems of control

**In humans the "normal' processes of control of attention can be disrupted or modified in a variety of ways.**

- **By external distractors, e.g. a loud noise or bright flash.**
- **By rival interests: not finishing one task because you remember another you want to complete.**
- **By strong emotional states (e.g. grief, anger, excited anticipation) that cause attention to be diverted from explicitly selected objects or tasks.**
- **By pathological states in which attention control mechanisms malfunction. (E.g. ADHD)**

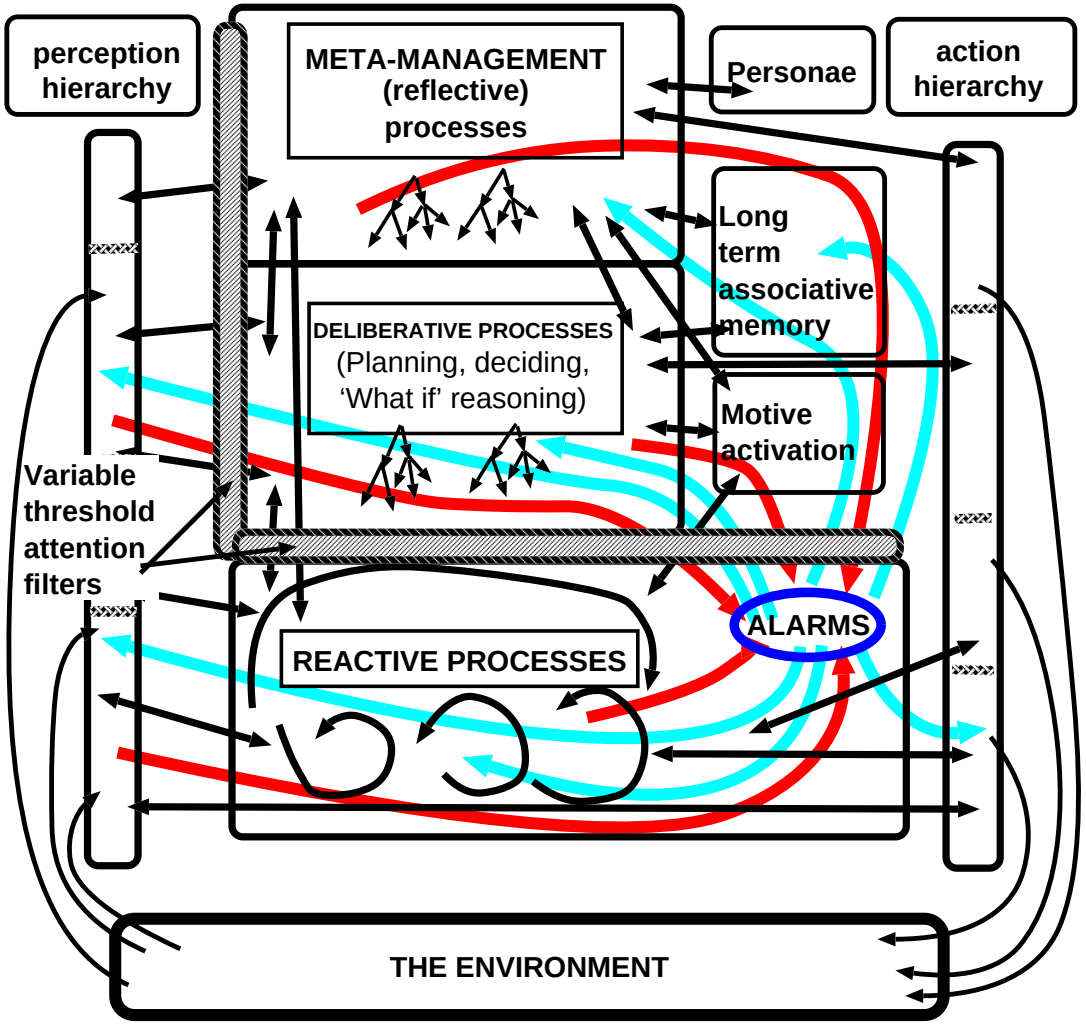
**Understanding the varieties of control and loss of control will help us understand more about normal mental states, such as emotional states, and may also help us predict a variety of architecture-related pathologies.**

**The neuropsychiatrist Russell A Barkley has written an important book (ADHD and the nature of self-control, The Guildford Press, 1997) which combines deep theorising (partly inspired by Bronowski) with a lot of empirical evidence. I believe this work sheds light on requirements for meta-management systems concerned with high level control functions. He seems to be completely unaware of work in AI, though there is a brief reference to computer models.**

# H-Cogaff – a first draft model of your mind

**The Birmingham "CogAff" project has been developing a framework for characterising a wide variety of types of minds, of humans, other animals, and possible future robots.**

**The framework incorporates evolutionarily ancient mechanisms co-existing and co-operating or competing with new mechanisms capable of doing different tasks (e.g. reasoning about what might happen). The figure gives an "impressionistic" overview of some of the complexity in our first draft H-CogAff architecture.**




**E.g. different sorts of emotions are generated in different levels.**

**More details including papers, slide presentations and software tools can be found at http://www.cs.bham.ac.uk/research/cogaff**

**Will this be useful in designing our robot children?**

# Applications of a working system

## Suppose we manage to produce such robots. What then?

**Imagine giving a working version of such a system to a student of developmental psychology to play with, including providing the opportunity both to look inside the thinking processes of the learner and to tinker with them.**

**Imagine giving a working version of such a system to a child, whose task is to help the robot overcome obstacles and difficulties, by giving it explanations of what it is doing wrong and descriptions of what it needs to do, or merely hints. Contrast programming the robot – as was done in the 1970s when children tried to get the LOGO turtle to make pictures.**

**Imagine a variant of the problem where one of the robots has some sort of emotional problem, e.g. gets very angry when things don't work, and also gets angry when anyone tries to help. Could this be useful in some contexts where counsellors, therapists, or teachers are being trained? (Or "difficult" children?)**

**CONJECTURE: we are talking about a very powerful kind of educational environment. It's not like the computer-based scenario-generators that just enable a child to program synthetic characters to say things and move around: for those don't have any notion of success or failure – anything goes.**

# Intelligent display manager

Insofar as there is a 2-D graphical display, and the simulated 3-D domain is fairly complex, there will be different possible views of what is going on, all involving projections from 3-D to 2-D but giving different information about the scene.

In Virtual Reality display systems analog controls are often provided to allow the user's viewpoint to move around, with variations like location, height, tilt, zoom, fish-eye view, etc.

Our 'manipulable world' scenario can be extended by allowing a user at the terminal to communicate with the display manager in a fashion that does not require learning to control such analog interfaces.

Give me a window showing the robot from the other side.

New window appears

I can't see what the hand is grasping.

Angle of view is altered, to show the whole of the hand

Fine, please zoom in a bit, ... further ... that's it.

Hmm, ... I'd like to see the clearance between the hand and the machinery.

Display manager: clearance above, below, or on one side?

Clearance on the left side (of the display).

This sort of thing could be useful not only for controlling complex machinery, but also in interactive instruction manuals, and of course many kinds of computer entertainments.

# Can't it all be evolved, or learnt?

**Naturally intelligent systems (ants, aardvarks, apes, and advocates of self-organising systems) are all existence proofs of the possibility of intelligent agents that are not designed or programmed. So can't we just evolve these robots?**

**Yes if you have 4,000,000,000 years, and a large planet on which to do your evolutionary experiments.**

**Why not start them off with no knowledge and let them learn or be trained, so that we don't have to program them?**

**It's true that we don't have to program baby humans or baby bonobos. But the fact that infants of different species can learn different things indicates that they start off with different learning engines provided by evolution. No amount of training will teach a bonobo, or a chicken, transfinite set theory, for instance, or even how to write java programs.**

**For a discussion of some of the innate conceptual apparatus and ontologies required for a human-like new-born learner see talk 14 here:**

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

**(slides attacking the notion of symbol-grounding and philosophical "concept-empiricism".).**

**QUESTION: How much knowledge about space, time, motion and causality has to be programmed in from birth? What about knowledge of mind?**

# What sort of knowledge of numbers?

**In *The Computer Revolution in Philosophy*, chapter 8, I discussed a certain stage in learning about numbers (child about 5 years old):**

**Adult: Can you count up to twenty?**
**Child: One two three four five six seven eight nine ten eleven twelve thirteen fourteen fifteen seventeen eighteen twenty.**
**A: What comes after three?**
**C: One two three four — four.**
**A: What comes after eight?**
**C: Four**
**A: What comes after six?**
**C: Don't know**
**A: What comes before two?**
**C: One**
**A: What comes before four?**
**C: Five**
**A: How many fingers on my hand?**
**C (counting fingers): One two three four five**
**A: What's two and three?**
**C (counting fingers): One two three four five. Five.**

**The chapter discusses association structures and the operations on those structures that a child might learn, e.g. in order to be able to answer more questions or to count backwards, etc. Could one of our robots do all that?**
**http://www.cs.bham.ac.uk/research/cogaff/crp/chap8.html**

# Why are humans different?

**Conjecture:**

**Human linguistic abilities, mathematical abilities, artistic abilities, not found in other animals depend on architectural differences that are mainly innately determined even if the architectures are not fully formed at birth. (Compare the distinction between precocial and altricial species.) Architectural features include:**

- **Having all three sorts of layers: reactive, deliberative and meta-management (reflective)**
- **Having particular forms of re-usable short term memory enabling construction of temporary representations with varying structures and complexity**
- **Having particular forms of representation that permit the construction of complex descriptions, including descriptions of relations between descriptions,**
- **Having mechanisms for creating matching and modifying such descriptions.**
- **Having storage mechanisms that allow such descriptions to be made available for long term use, with flexible context-sensitive retrievable mechanisms**
- **Having representational mechanisms that allow new ontologies to be constructed including components that are not all *explicitly definable* in terms of previous concepts. (Compare Carnap on 'Meaning postulates'.)**
- **Allowing different dedicated memory modules to be constructed that store related items (e.g. face memory, lexical memory, grammatical memory, plan memory, memory connected with a particular sort of game, musical memory).**

**I suspect we now understand a lot less than 1% of the mechanisms.**

# How to learn from children

## Watch a baby a few months old as a bottle moves towards its mouth.

There's a huge amount of more or less coordinated activity: arms and legs waving, mouth opening, tongue moving out and down, eyes converging on the bottle.

Many questions arise:

- Does the baby have any knowledge of what is going to happen: e.g. the nipple will eventually be fully inserted so that sucking can begin?
- Does the baby have any notion of what it is doing, or is it still just a bundle of reactions trying to get themselves woven into effective behaviour sequences?
- Is this an *essential* component of an infant acquiring a grasp of concepts relating to space, motion and action, or could its understanding grow in some other way?
- Do all those experiments done on newborn infants by developmental psychologists tell us anything about the architecture, representations, mechanisms being used, or are they all engaged in fanciful, undisciplined interpretation of highly ambiguous data (like many opinionated archaeologists)?

Compare other nascent competences: e.g. an infant lying on its belly, unable to crawl, yet apparently trying. Does it really know what it cannot yet do and want to do it? Or is something else going on?

Maybe we need to design various kinds of simulated infant learners, to find out what exactly needs to be explained.

# Learning to see new possibilities

I watched a young child struggling to open a tin containing his blocks: it was a large catering-size coffee tin, with a flanged lid. He had previously used a long thin screwdriver to lever off the lid, but he could not find it.

Suddenly (like Kohler's apes) he had an idea, grabbed the lid of a smaller tin that was close by and used that circular lid instead of a screwdriver.

Apparently he had seen a deep structural relationship between the missing screwdriver and the circular lid. I wonder if we have any language to express precisely what he saw.

Did he already have the ontology required, and merely notice a new way of applying it, or did he have to extend his ontology so that the screwdriver and the circular lid fell under some new concepts, which he had not previously used?

How does an ontology grow? Where and how is the ontology represented in the child?

Does the child understand the causal connections between various actions and their consequences, as it goes through the actions need to prize open the lid?

What is it to understand causal connections?

How much of this will have to go into our simulated robots, if they are to put everything together again?

For more examples see the 1960s books of John Holt, e.g. *How children learn*.

# Visual architectures

**In Chapter 9 of *The Computer Revolution in Philosophy*, chapter 8, I discussed how vision (like other forms of perception?) might involve several concurrent layers of processing. http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html**

**That was justified in terms of examples of messy pictures that required a mixture of concurrent top-down and bottom up processing.**

**At that stage I had not grasped the need for concurrent layers feeding into different parts of a central architecture: reactive mechanisms, deliberative mechanisms, reflective (meta-management) mechanisms.**

**(Now called 'multi-window' perception as contrasted with 'peephole' perception.)**

**Can we demonstrate the need for and power of this within the manipulable world project without having to build physical robots: maybe some aspects could be simulated, with enough detail to illustrate the main architectural points – followed later by implementation in physical robots as an existence proof, or something?**

**Eventually we'll need to do this with physical robots with real cameras, and also other sensors that can feed into multi-layered perceptual mechanisms.**

**(Likewise multi-layered control mechanisms coordinating intricate collections of motors (e.g. muscles) producing many kinds of motion to be integrated in effective actions.)**

# Specify criteria for evaluating progress.

**At least three sorts of criteria**

- **Answering hard scientific questions**
  **(producing explanatory theories of mind)**
- **Producing applications**
  **(things people or companies want to buy)**
  - **games and entertainments (including new synthetic agents to take part in existing computer games, e.g. “Unreal tournament”).**
  - **personal assistants**
  - **useful new intelligent robots (e.g. robot nurses – Takeo Kagami)**
  - **tutoring systems**
  - **perhaps using an extendable “shell” that accepts plugins.**
- **Applying the knowledge we have gained in other ways than by using the system to perform some tasks:**
  - **Teaching psychologists and neuroscientists by giving them models to play with and modify**
  - **Designing new forms of therapy, teaching, counselling, based on our improved understanding of human mental functioning, learning, etc.**
  - **....**

# But will the robots really understand? Or really think, feel, desire, plan, learn, feel pain, ...?

**What's the difference between *understanding* and *really understanding*?**

**Philosophers have debated this for years. Many engineers don't care whether their robots "really" have mental processes as long as they produce the right behaviour. Scientists are often ambivalent: and some even think the answer depends on finding neural correlates of mental processes and replicating them.**

**Our approach:**

- **If anyone can identify a kind of behavioural competence that is missing from our robots and which ought to be there (to meet scientific or practical requirements) then we can accept that as a challenge to extend our designs.**
- **If anyone can clearly identify and describe some feature of human internal processes (thinking, perceiving, learning, experiencing) that is missing from the virtual machine processes in our robots, then we can accept that as a challenge to extend our designs.**
- **If anyone argues that it is not enough to replicate external behaviours and internal virtual machine states and processes, along with their causal relationships, then if he/she can specify any clear additional requirement we'll pay attention. Often such additional requirements evaporate into incoherent internal ostensive definitions: "*This* is missing"..**

**(See talk 9 on varieties of consciousness here: http://www.cs.bham.ac.uk/~axs/misc/talks/)**

# The importance of scenarios

**Well-described scenarios constitute a medium for posing challenges.**

**If we claim to have explained or modelled some kind of capability or some class of mental states or processes, then a perceptive critic may be able to dream up an interesting, challenging, scenario on which the design can be tested.**

- **B sees some action done by A, and then makes an inference about A's beliefs, or intentions, or feelings. (This could benefit from AI work on plan inference.)**
- **Aetiology of emotions: e.g. A does something that makes B irritated, then increasingly angry until B explodes with anger. The anger then subsides, B starts to feel guilty and then wants to apologise to A.**
- **Social "contractual" relations: A promises to do something for B, then fails to do it, and offers no explanation. B feels indignant. A is unrepentant. B starts threatening revenge. A starts feeling worried....**
- **Robots go through mental processes that have nothing to do with current tasks. (See Erik Mueller's book 'Daydreaming'.)**
- **A robot gets interested in a mathematical problem, searches for a proof, becomes elated when one seems to work, becomes worried about whether it really works, thinks hard and finds a counter-example, becomes depressed and starts thinking about something else, later remembers the proof and thinks of a way to deal with the counter-example, is elated again, etc. All without external behaviour.**

# Meta-scenarios

If something can happen to a robot, or happen in a robot, then the robot may be able to talk about it to others, e.g. answering questions, or expressing puzzlement.

I.e. for every scenario in which process P occurs, we can construct additional scenarios in which the occurrence of P is talked about, thought about, wondered about, etc.

These are meta-scenarios.

Meta-scenarios can challenge different kinds of features in the robots from the original scenarios.

Sometimes it is easier to implement the ability to do X, than the ability talk or think about someone doing X. (Why?)

Sometimes it is harder. (Why?)

To help guide the design process we need something like a taxonomy of types of scenarios, so that we can fairly easily construct more and more challenging ones, to drive the development process.

# Towards specification of the tools

**Designing a multi-competence robot's mind will need special tools, including tools for building actual or simulated physical robots and their environments, and tools for building their minds.**

- Maybe we can use some off-the-shelf simulation software to run the physical simulation and generate graphics.
- One option is to interact with a game engine (several AI researchers are designing AI agents that communicate with a game server).
- But game engines may not support the kinds of articulated movable objects that we need.
- Will the ODE system suffice (at least initially)?
  http://www.q12.org/ode/ode.html
- Can something be done with SodaPlay beasts and machines?
  http://www.sodaplay.com

**The much harder task will be building the minds of the robots, even if we build minds that mostly build themselves. This will require tools of a different sort, with quite different problems of designing, implementing, testing and debugging.**

E.g. physical simulations can be tested by looking at graphical displays.

How do you test and debug very complex information processing in virtual machines: externally visible behaviour will rarely give enough information.

# We'll need a collection of tools and protocols for developing, testing, and running

- Tools for developing rule-based systems
- Tools for developing neural nets
- OOP tools for developing re-usable classes and methods
- Tools for combining mechanisms to operate concurrently within an agent architecture (including relative speed variations).
- Tools and protocols for distributing agents, or agent components across a collection of CPUs.
- Tools for developing graphical interfaces – for development and for interacting with running systems.
- Tools for developing natural language interfaces (with textual and spoken language).
- Tools to support shared development and delivery of re-usable components.
- Tools to support debugging and testing.

**Start by identifying and comparing some existing toolkits and developing criteria for selection. E.g. explain why so many people like using the "Genera" system (Lisp machine environment).**

# Some example tools/toolkits

- CYC
- Architecture specific toolkits, e.g. Soar, Act-RP, Jack, ...
- Architecture-neutral agent toolkits SimAgent, Mozart, ...
- Tools for supporting remote shared-source collaborations (as in GNU, Mozilla, OpenOffice)

## Contrast

– tools that are good for novice users to try out and learn about agent development

– tools that are good for extended use by expert engineers

**Will our SimAgent toolkit be useful for implementing minds of first draft agents?**

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

**It is**

– Architecture neutral
– Difficult for beginners though useful for experts
– Free of charge and open source
– Based on a language whose power and flexibility is comparable to Common Lisp, though some people find it easier to read and use.

# To be continued

## Lots more to be said

- about the world
- about the agent architectures and mechanisms
- about motivation, emotions, and other affective states and processes
- about the scenarios to be achieved
- about how to build on existing work (e.g. CYC? simulation engines, game engines, NLP engines, work on vision, planning, learning, problem solving, robot control, neural nets...)
- about long term extensions and the module ‘plug-in’ capabilities.
- about project management and evaluation

# Some links – in need of organisation

- Freddy the versatile robot, Edinburgh 1973:
  http://www-robotics.cs.umass.edu/~pop/VAP.html
- The AAAI web site on AI http://www.aaai.org/aitopics/
- RoboCup Rescue web page http://robomec.cs.kobe-u.ac.jp/robocup-rescue/
- The RoboCup federation http://www.robocup.org
- Minsky's home page http://www.media.mit.edu/~minsky/
- McCarthy's home page http://www-formal.stanford.edu/jmc/
- Nebel's presentation on RoboCup
  http://www.informatik.uni-freiburg.de/~nebel/ACAI-01/ppframe.htm
- Forbus' group (papers on qualitative physics, common sense reasoning, analogy and similarity, sketching, educational software, interactive entertainments, etc.
  http://www.qrg.ils.nwu.edu/papers/papers.htm
- Ben Kuipers' group http://www.cs.utexas.edu/users/qr/
- Tony Cohn's group http://www.comp.leeds.ac.uk/qsr/ ;; Tony Cohn's group, focuses on
- OpenCyc http://www.opencyc.org/
- The Birmingham CogAff project http://www.cs.bham.ac.uk/~axs/cogaff.html
- The Birmingham ATT-Meta project http://www.cs.bham.ac.uk/ jab/ATT-Meta/

# LOTS MORE SHOULD BE HERE...