Invited Presentation at Public Session of AISB’08, 3rd April 2008, Aberdeen

http://www.aisb.org.uk/convention/aisb08/

What designers of artificial companions need to understand about biological ones.

Aaron Sloman

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

These slides (expanded since the event) are in my ‘talks’ directory:

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#aisb08

Apologies

I apologise

- For slides that are too cluttered: I write my slides so that they can be read by people who did not attend the presentation.
- So please ignore what's on the screen unless I draw attention to something.

No apologies

For using linux and latex

This file is PDF produced by pdflatex.

NOTE:

To remind me to skip some slides during the presentation they are ‘greyed’.

This presentation, in Aberdeen in April 2008 overlaps with a presentation at a meeting on Artificial/Digital companions at the Oxford Internet Institute in October 2007, and several other presentations in this directory:

```
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/
```

Plan – may not have enough time

This is what I aim to do

- Present some examples (using videos) of biological competences in children and other animals.
- Discuss some of the requirements for artificial companions (ACs).
- Distinguish relatively easy goals and much harder, but more important, goals for designers of ACs (Type 1 and Type 2 goals, below).
- Assert that current techniques will not lead to achieving the hard goals.
- Suggest ways of moving towards achieving the hard goals, by looking more closely at features of how the competences develop in humans.
- In particular there is a lot of knowledge about the space and structures and processes in space that a young child develops (some of it shared with some other animals) that forms part of the basis of a lot of later learning.
- No AI systems are anywhere near the competences of young children.
- Trying to go directly to systems with adult competences will produce systems that are either very restricted, or very brittle and unreliable (or both).

Show some videos of biological competences

Parrot scratching its neck with a feather.

Felix Warneken's videos

`http://email.eva.mpg.de/~warneken/`
Chimp (Alexandra) and fallen lid
Child and cabinet

Warneken was trying to show that both chimpanzees and pre-verbal human toddlers could be altruistic – i.e. motivated to help others.

My interest is in the forms of representation and architectures that allow such individuals to represent complex structures and processes in the environment and to detect that someone else has a specific goal that is not being achieved.
Forming the motive to help with that goal, and acting on it, must build on sophisticated cognitive competences.

Child playing with trains

a 30 month old child sitting on the floor playing with trains has a rich store of usable information about what is going on around him, including some of the processes and relationships that he is no longer looking at. Video included here:
`http://www.cs.bham.ac.uk/research/projects/cosy/conferences/mofm-paris-07/sloman/vid`

Child jumping?

A 27 month old child stepping into his father's shoes, then jumping/walking around in them (with music in the background). What forms of representation, concepts and knowledge must he be using?

Betty (New Caledonian crow) makes hooks out of wire.

`http://users.ox.ac.uk/~kgroup/tools/crow_photos.shtml`

Example of your own spatial competence

Have you ever seen someone shut a door by grasping the edge of the door and pushing or pulling it shut?

Would you recommend that strategy?

What will happen to someone who tries?

How do you know?

Humans, and apparently some other animals, have the ability to represent and reason about a wide class of spatial structures and processes – including some that have never been perceived.

This ability develops in different ways in different environments and in different individuals – not all cultures have hinged doors.

What processes in human brains, or in computers, could enable such visual reasoning to occur?

I'll suggest that such capabilities are likely to be required in certain kinds of Artificial Companions looking after people living in their own homes.

What are designers of artificial companions trying to do?

Common assumptions made by researchers:

- ACs will be able to get to know their owners and help them achieve their goals or satisfy their needs.
- Their owners could be any combination of: elderly, physically disabled, cognitively impaired, lonely, keen on independence, ...
- ACs could provide assistance using pre-built knowledge plus knowledge gained from external sources, including central specialist databases that are regularly updated, and the internet.
- They may help with practical problems around the home, as well as providing help, warnings and information requested.
- They could also provide company and companionship.

My claim:

> The detailed requirements for ACs to meet such specifications are not at all obvious, and will be found to have implications that make the design task very difficult, in ways that have not been noticed. Achieving the goals may not impossible provided that we analyse the problems properly.

Some people argue that new kinds of information-processing machines will be required in order to match what biological systems can do.

As far as I am concerned that is an open question: dogmatic and wishful answers should be avoided.

Why will it be so hard?

Because:

(a) Motivations for doing these things can vary.

(b) The more ambitious motivations have hidden presuppositions.

(c) Those presuppositions concern the still barely understood capabilities of biological companions.

(d) Getting those capabilities into machines is still a long way off:

Mainly:

Artificial Companions will have very open-ended, changing sets of requirements, compared with domain-specific narrowly targeted AI systems

Examples of domain-specific, task bounded systems:

a flight booking system
route advising system for a fixed bus or train network
a theorem prover or proof checker to help someone working in group theory
a de-bugger for student programs for producing a certain class of graphical displays

In humans new skills or knowledge can usually be combined with old skills and knowledge in creative ways: a Type 2 AC (explained below) will need to do that.

Most AI systems are not able to support such extendability and flexibility.

What sorts of designs could support that?

We are not even close.

We are nowhere near getting machines to have the capabilities of young children or nest-building birds or young hunting mammals or ...

There are some impressive (often insect-like) AI systems that have been highly trained or specially designed for specific tasks, but many insects are far more sophisticated.

Such AI systems do not know what they are doing or why, or what difference it would make if they did something different.

However, AI has achieved some very impressive results, e.g. chess programs that beat grand-masters, mathematical theorem provers, planning systems, and search engines like google, which in their narrow fields completely outstrip human intelligence.

Even when they are spectacularly good in their own domains, such systems are not designed to be capable of being combined with other competences to match human creativity and flexibility: they may scale up but they don't scale out.

A world-beating chess program cannot decide how to play with a young enthusiastic but inexpert player so as to encourage and inspire that child to go on learning.

A language-understanding program typically cannot use visual input or a gesture to help it disambiguate an utterance.

A vision program typically cannot use linguistic input to guide its understanding of a complex scene or picture.

Scaling up requires architectures, forms of representation and mechanisms that allow different kinds of knowledge or skill to be combined opportunistically.

I am not saying it is impossible

I am not saying it is **impossible** for AI to achieve these things:

but first the researchers have to want to achieve them,

and then they have study the natural systems to find out a lot more about the variety of **tasks** and ways of getting things right or getting them wrong,

and ways of debugging the faulty solutions

I.e. they need to study different combinations of **requirements**

instead of just focusing on **mechanisms** and trying to make them work a bit better.

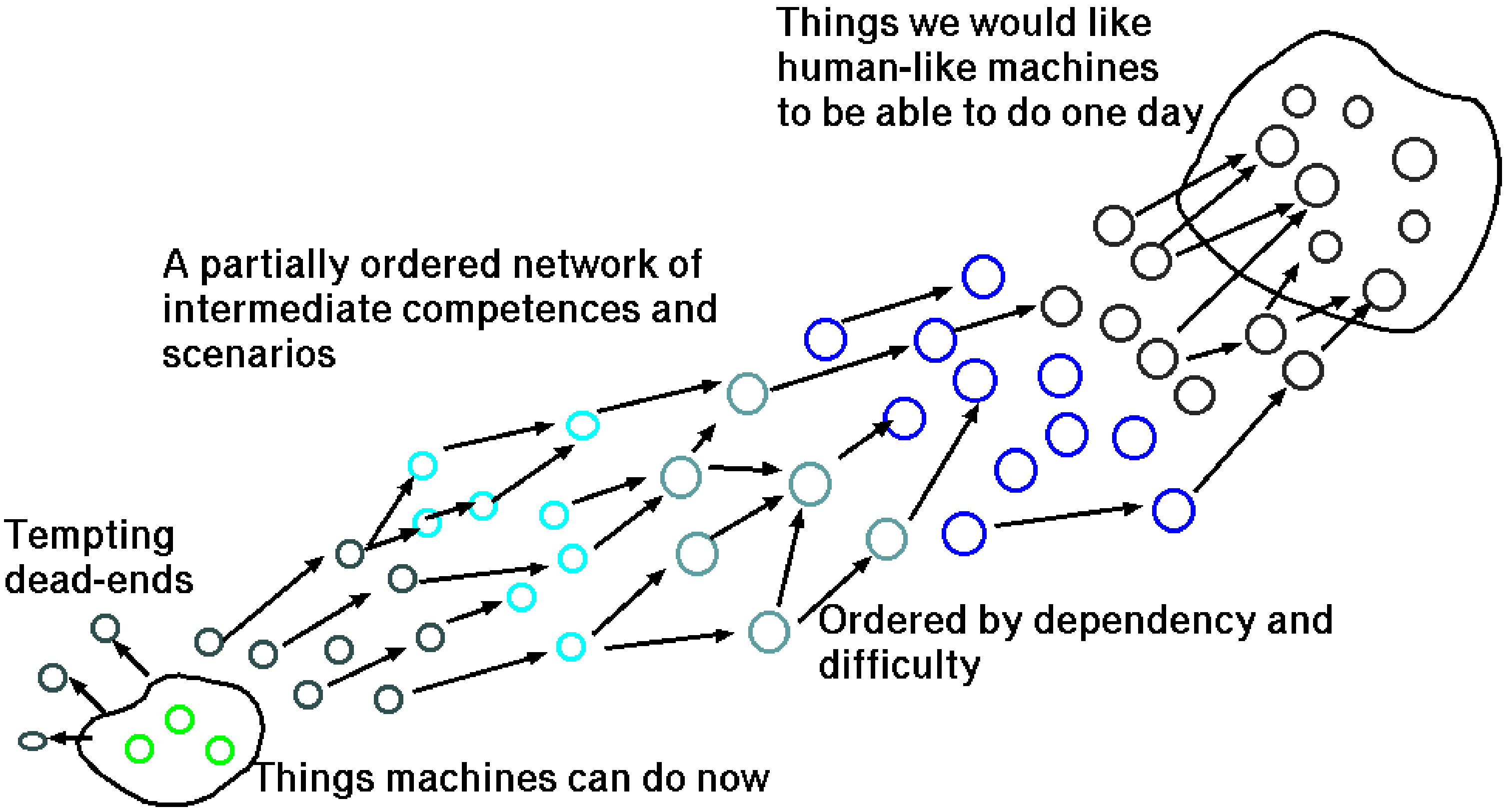
It is specially important not to use fixed narrowly-focused benchmark tests to drive research: all that will produce is (at best) systems that are good on those benchmarks.

Benchmarks can be very useful for engineering projects.

Scientific AI research has been seriously harmed by benchmarks.

Beware the stunted roadmap trap

The most obvious and tempting research roadmaps may lead to dead-ends.


Things we would like
human-like machines
to be able to do one day
A partially ordered network of
intermediate competences and
scenarios
Tempting
dead-ends
Ordered by dependency and
difficulty
Things machines can do now

People who don’t do the right sort of requirements-analysis see only the routes to “tempting dead ends” – the stunted roadmap: but they don’t recognise them as dead ends because those routes lead to ways of improving their programs – to make them a little better.

Type 1 goals for ACs

People working on artificial companions may be trying to produce something that is intended to provide one or more of the following kinds of function:

Type 1 goals:

Engaging functions

- TOYS:
  a toy or entertainment for occasional use (compare computer games, music CDs)

- ENGAGERS:
  something engaging that is used regularly to provide interest, rich and deep enjoyment or a feeling of companionship (compare pets, musical instruments)

- DUMMY-HUMANS (digital pacifiers?):
  something engaging that is regarded by the user as being like another caring, feeling individual with which a long term relationship can develop – even if it is based on an illusion because the machine is capable only of shallow manifestations of humanity: e.g. learnt behaviour rules, such as nodding, smiling, head turning, gazing at faces, making cute noises, or in some cases apparently requesting being comforted, etc.

These goals don’t interest me much, though I have nothing against people working on them, provided that they are not simply pursued cynically for financial gain.

Type 2 goals for ACs

Some people working on artificial companions hope eventually to produce something that is intended to provide one or more of the following kinds of function:

Type 2 goals:

Enabling functions

- HELPERS:

  ACs that can reliably provide help and advice that meets a fixed set of practical needs that users are likely to have, e.g.:

  Detecting that the user has fallen and is not moving, then calling an ambulance service.

- DEVELOPING HELPERS:

  a helper that is capable over time of developing a human-like understanding of the user's environment, needs, preferences, values, knowledge, capabilities, e.g.:

  Noticing that the user is becoming more absent-minded and therefore learning to detect when tasks are unfinished, when the user needs a reminder or warning, etc.

  More advanced competences of this sort will require the next category of AC.

- CARING DEVELOPING HELPERS:

  a developing helper that grows to care about the user and really wants to help when things go wrong or risk going wrong, and wants to find out how best to achieve these goals.

  Caring will be required to ensure that the AC is motivated to find out what the user wants and needs and deals with conflicting criteria in ways that serve the user's interests.

I am more interested in these, but they are very hard.

Why out of reach?

Current AI is nowhere near meeting the following requirements:

- An AC may need to understand the physical environment
  - what can't you easily do with a damaged left/right thumb or a stiff back, etc..
  - noticing a dangerous situation, e.g. apron string looped round pan handle,
  - coffee spilt on electric kettle base, blows a household fuse:
    when will it be safe to turn on the kettle again?
  - helping a user solve the problem of losing balance when soaping a foot in a shower,

- An AC will need to understand human minds – at least as well as humans do

  E.g. guessing how a patient is likely to react to a variety of new situations.

  Knowing what a patient is likely to find hard to understand.

  Knowing how detailed explanations should be.

  Diagnosing sources of puzzlement, when the patient is unable to understand something.

  Designing machines with these competences will require us to have a much deeper understanding of the information processing architectures and mechanisms used by humans who do have those competences.
  Trying to use statistical learning to short-cut that research will produce shallow, limited, and fragile systems.

  I would not trust them with an elderly relative of mine.

- We need progress on architectures that allow a machine to develop new motives, preferencs, ideals, attitudes, values ....

Developing new motives on the basis of deep knowledge and real caring

The AC need will need to derive new caring motives on the basis of those competences:

Starting from real desire to help, it will need to be able to work out new ways to help, or to avoid being unhelpful, in different situations.

It may change some of its motives or its preferences if it discovers that acting on them leads to consquences it alread prefers to avoid, e.g. the user being unhappy or angry.

The AC should want to act on the conclusions based on new motives and preferences.

This has ethical implications, mentioned later.

The requirement for future robots to be able to develop their own motives was discussed in *The Computer Revolution in Philosophy* (1978), e.g. in Chapter 10 and the Epilogue
`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`

How can we make progress?

Since humans provide an existence proof, one way to meet these objectives in a reliable fashion will be to understand and replicate some of the generic capabilities of a typical young human child, including the ability to want to help.

Then other capabilities can be built on top of those

(layer after layer, using the ability of the architecture to extend itself).

A human child does not start with a “tabula rasa” (blank slate): the millions of years of evolution that produced human brains prepared us to learn and develop in specific ways in specific sorts of environments.

Adult human capabilities (including helping, caring, learning capabilities) develop on the basis of a very rich biological genetic heritage, plus interaction with the environment over many years,

though not necessarily in a normal human body:

Compare humans born limbless, or blind, or deaf, or with cerebral palsy,
or Alison Lapper. (See `http://www.alisonlapper.com/`)

The brains of such people matter more for their development than their bodies.

But their brains are products of evolution – evolution in the context of normal bodies.

Note: Humans are also an existence proof that systems that develop in these ways can go badly wrong and do serious harm, so many precautions will be needed.

Limitations of current methods

Explicitly programmed systems are very limited.

It is not necessarily so, but in practice the difficulties of hand-coding are enormous.

Compare the number of distinct design decisions that go into producing a Boeing 747 or an Airbus 380.

Could a similar amount of human effort produce a worthwhile AC?
Only if the **design** were based on a deep understanding of **requirements**.

Designers try to get round limitations by producing systems that learn.

But the learning methods available so far are far too impoverished

Many of them merely collect statistics based on training examples, e.g. learning what to do when, and then use the statistics to modify their behaviour so as to conform to those statistics.

Such statistical learning methods with little prior knowledge about at least the **type** of environment in which learning will occur, will always be limited, fragile and unreliable (except for very restricted applications).

See John McCarthy's "The Well-designed Child"

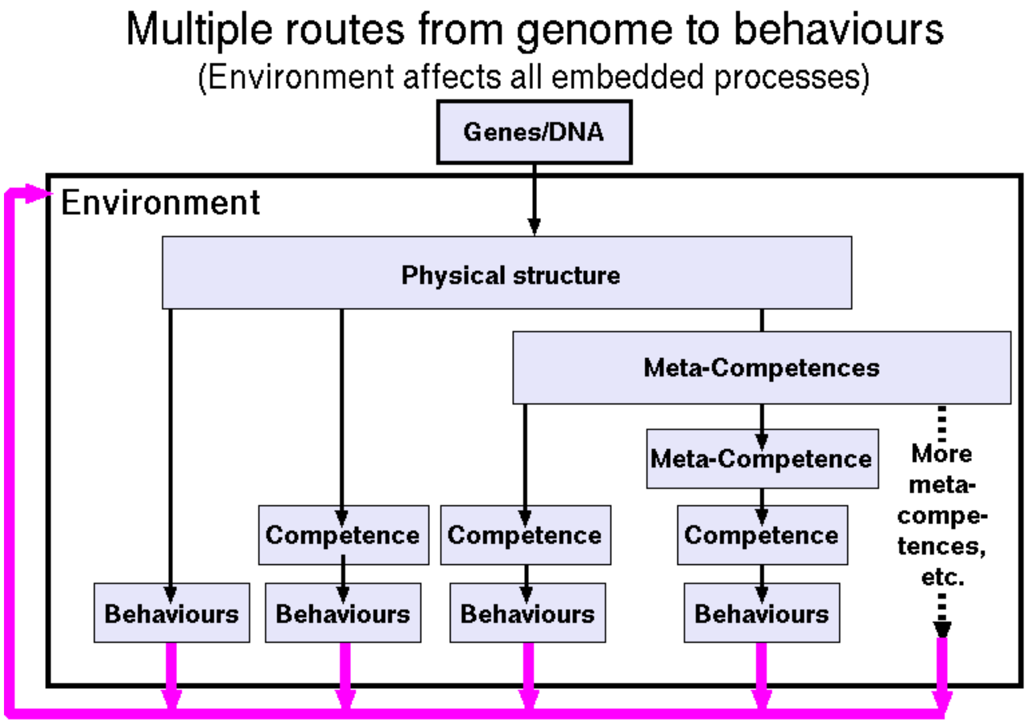
> "Evolution solved a different problem than
> that of starting a baby with no a priori assumptions."

On his web page.

`http://www-formal.stanford.edu/jmc/child.html` (Also *AI Journal* December 2008)

A more structured view of nature-nurture trade-offs

Work done mainly with Jackie Chappell, published in IJUC 2007.
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609`



(Chris Miall helped with the diagram.)

Routes on left from genome to behaviour are more direct.
(The vast majority of organisms have no other option.)

Routes towards the right may involve many different layers of competence, developed in stages, with different ontologies, based partly on learning new ways to learn.

See slide below on an architecture that builds itself.

Example: A kitchen mishap

Many of the things that crop up will concern physical objects and physical problems.

Someone knocks over a nearly full coffee filter close to the base of a cordless kettle.
This causes the residual current device in the central fuse box to trip, removing power from many devices in the house.

You may know what to do: unplug the base, reset the RCD, and thereby restore power.

But if you not sure whether it was safe to use the kettle after draining the base, and when you try it later the RCD trips again, leaving you wondering whether it would ever be safe to try again, or whether you should buy a new kettle.

Would the AC be able say: "Try opening the base, dry it thoroughly, then replace it, and the kettle can be used again"?

- Should an AC be able to give helpful advice in such a situation?
- Would linguistic interaction suffice? How?
- Will cameras and visual capabilities be provided?

What visual capabilities: current AI vision systems are close to useless!

Canned responses – and intelligent responses

- The spilt coffee was just one example among a vast array of possibilities that will vary from culture to culture, from household to household within a culture and from time to time in any household, as the people and things in the house change.

- Of course, if the designer anticipates such accidents, the AC will be able to ask a few questions and spew out relevant canned advice, and even diagrams showing how to open and dry out the flooded base.

- But suppose designers had not had that foresight: What would enable the AC to give sensible advice?

How can we hope to produce an AC with something like human varieties of competence and creativity?

- If the AC knew about electricity and was able to visualise the consequences of liquid pouring over the kettle base, it might be able to use a mixture of geometric and logical reasoning creatively to reach the right conclusions.

- It would need to know about and be able to reason about spatial structures and the behaviour of liquids.

  This will require us to go far beyond the 'Naive Physics' project

  - Although Pat Hayes described the 'Naive physics' project decades ago, it has proved extremely difficult to give machines the kind of intuitive understanding required for creative problem-solving in novel physical situations.
  - In part that is because we do not yet understand the forms of representation humans (and other animals) use for that sort of reasoning
  - The Naive Physics project used logic to formalise everything, and the only use of the knowledge was to perform deductions, whereas humans and other animals use a formalism that allows the knowledge to be used in perception, action and intuitive causal reasoning.

  See these papers and presentations on understanding causation with Jackie Chappell:
  `http://www.cs.bham.ac.uk/research/projects/cogaff/talks#wonac`

Understanding shower-room affordances

Understanding a human need and seeing what is and is not relevant to meeting that need may require creative recombination of prior knowledge and competences.

Suppose an elderly user finds it difficult to keep his balance in the shower when soaping his feet, and asks the AC for advice.

He prefers taking showers to taking baths, partly because showers are cheaper.

How should the AC react on hearing the problem?

Should it argue for the benefits of baths?

Should it send out a query to its central knowledge base asking how how people should keep their balance when washing their feet?

(It might get a pointer to a school for trapeze artists.)

The AC could start an investigation into local suppliers of shower seats: but that requires working out that a seat in the shower would solve the problem, despite the fact that showers typically do not and should not have seats.

What if the AC designer had not anticipated the problem?

What are the requirements for the AC to be able to invent the idea of a folding seat attached to the wall of the shower, that can be lowered temporarily to enable feet to be washed safely in a sitting position?

Alternatively what are the requirements for it to be able to pose a suitable query to a search engine?

How will it know that safety harnesses and handrails are not good solutions?

The ability to deal with open-ended sets of problems

I have presented a few examples of problems an AC may have to deal with around the house, all involving perceiving or reasoning about a novel problem situation.

The set of possible problems that can arise is enormously varied,

- partly because of the open-ended variability of ways in which physical objects (including humans, pet animals and inanimate objects) can interact
- partly because of the ways in which the contents of a house can change over time
- partly because of the ways in which a human's needs, preferences, desires, knowledge and capabilities can change over time

A **developing caring helper** will have to be able to deal with all that novelty.

Humans cope with novelty on the basis of very general capabilities for learning about new environments, which build on a their rich evolutionary heritage in ways that are not yet understood.

Likewise ACs will need to be able to 'scale out', i.e. to combine competences in new ways to deal with novel situations.

Is the solution statistical?

Can the required competences, including the ability to combine them creatively, be acquired by a machine trained on a very large corpus of examples in which it searches for statistically re-usable patterns.

The current dominant approach to developing language understanders, vision programs, robots, advice givers, and even some engaging interfaces, involves mining large collections of examples, using sophisticated statistical pattern extraction and matching, to find re-usable patterns.

This is sometimes easier than trying to develop a structure-based understander and reasoner with the same level of competence, and can give superficially successful results, depending on the size and variety of the corpus and the variety of tests.

But the method is inherently broken because, as sentences get longer, or semantic structures get more complex, or physical situations get more complex, the probability of encountering recorded examples close to them falls very rapidly to near zero – especially when problems include several modalities, e.g. perception, planning and language understanding.

A helper must use deep general knowledge to solve a novel problem creatively.

E.g. understanding a new utterance often requires using non-linguistic context to interpret some of the linguistic constructs.

How big is a pile of stones? It depends on the context. See:
http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0605

One of the problems of statistical/corpus-based approaches

An obvious problem, often forgotten in the flush of progress from 86% correct on some benchmark test to 86.35% is this

> however well a system trained on text does on text-based benchmark tests, it will not necessarily have any of the abilities that a human understander has to link the syntactic and semantic linguistic structures to the physical or world in which we perceive, move, manipulate and interact with things, including other language users.

Put another way: before a young child learns to understand and produce linguistic utterances she typically has rich physical and social interactions with things and people in the environment – and learning to talk is learning to use linguistic structures and processes to enrich those interactions.

Putting the understanding of a corpus based NLP system to the same tests a young human succeeds in would require linking the system to some sort of robot.

Some people are trying: that challenge is part of the specification of the EU Cognitive Systems initiatives in FP6 and FP7.

My impression is that progress is still very slow.

I think that is partly because the **problems** defining the **requirements** have not been studied in sufficient depth.

Size alone will not solve the problem.

There is no reason to believe that the fundamental problems will be overcome just by enlarging the corpus:

even a system trained not only on the whole internet but also on all human literature will not understand that the text refers to things that can be seen, eaten, assembled, designed, danced with, etc.

— unless it starts off with knowledge about the world somehow pre-compiled, and uses the textual analysis to abduce mappings between the textual structures and the non-textual world.

The myth of multi-modality

It is often assumed that the problem can be overcome by applying the statistical approach to a combination of modalities:

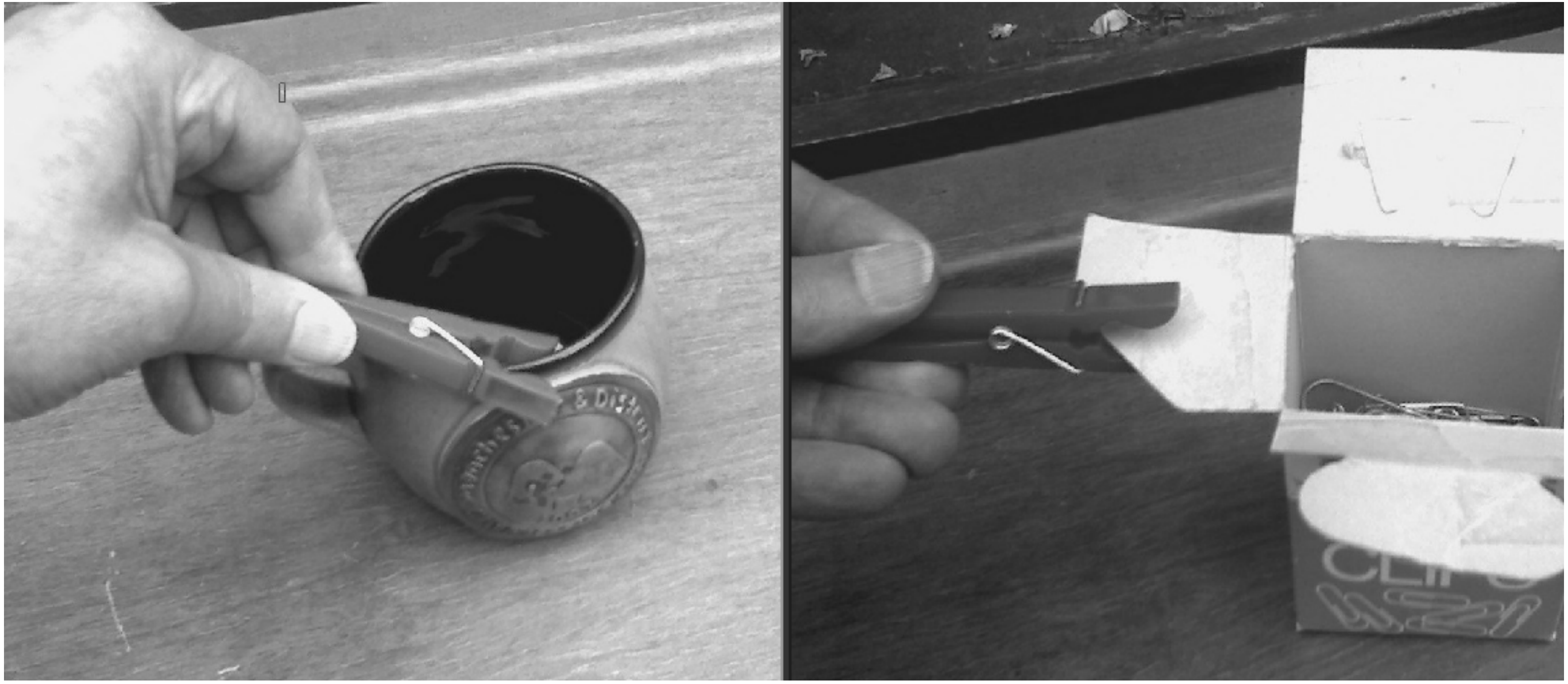
> give a system lots of images, or even videos, as well as audio-tracks, and maybe other kinds of sensory inputs, and then let it learn correlations between the things people see, smell, taste, feel, hear, etc. and the things they say or write.
>
> (That's the hope.)

But multimodality leads to a huge combinatorial problem because of the vast diversity of patterns of sensory and motor signals.

I conjecture that only a machine that is capable of developing an a-modal ontology, referring to things in the environment will be able to understand what people are talking about.

Consider the variety of sensor and motor signal patterns associated with the notion of something being grasped.

Grasping: Multiple views of a simple a-modal process of two 3-D surfaces coming together with another 3-D object held between them.



Four examples of grasping: two done by fingers, and two by a plastic clip. In all cases, two 3-D surfaces move together, causing something to be held between them, though retinal image projections and sensorimotor patterns in each case are very different.

You can probably see what is going on in these pictures even though they are probably very different in their 2-D structure from anything else you have ever scene.

But if you map them onto 3-D structures you may find something common, which cannot be defined in terms of 2-D image properties.

Why do statistics-based approaches work at all?

The behaviour of any intelligent system, or collection of individuals, will leave traces that may have re-usable features, and the larger the set the more re-usable items it is likely to contain – up to a point.

For instance it may not provide items relevant to new technological, or cultural developments or to highly improbable but perfectly possible physical configurations and processes.

So any such collection of traces will have limited uses, and going beyond those uses will require **something like the power of the system that generated the original behaviours.**

Some human expertise is like this.

How humans use statistical traces

In humans (and some other animals), there are skills that make use of deep generative competences whose application requires relatively slow, creative, problem solving, e.g. planning routes.

Frequent use of such competences trains powerful learning mechanisms that compile and store many partial solutions matched to specific contexts (environment and goals).

As that store of partial solutions (traces of past structure-creation) grows, it covers more everyday applications of the competence, and allows fast and fluent responses in more contexts and tasks.

A statistical AI system that cannot generate the data can infer those partial solutions from large amounts of data.

But because the result is just a collection of partial solutions it will always have severely bounded applicability compared with humans, and will not be extendable in the way human competences are.

If trained only on text it will have no comprehension of non-linguistic context.

Coping with novelty

The history of human science, technology and art shows that people are constantly creating new things – pushing the limits of their current achievements.

Dealing with novel problems and situations requires different mechanisms that support creative development of novel solutions.

(Many jokes depend on that.)

If the deeper, more general, slower, competence is not available when stored patterns are inadequate, wrong extrapolations can be made, inappropriate matches will not be recognised, new situations cannot be dealt with properly and further learning will be very limited, or at least very slow.

In humans, and probably some other animals, the two systems work together to provide a combination of fluency and generality. (Not just in linguistic competence, but in many other domains.)

Where does the human power come from?

Before human toddlers learn to talk they have already acquired deep, reusable structural information about their environment and about how people work.

They cannot talk but they can see, plan, be puzzled, want things, and act purposefully: They have something to communicate about.

E.g. see Warneken's videos.

That pre-linguistic competence grows faster with the aid of language, but must be based on a prior, internal, formal 'linguistic' competence

using forms of representation with structural variability and (context-sensitive) compositional semantics.

This enables them to learn any human language and to develop in many cultures.

Jackie Chappell and I have been calling these internal forms of representation Generalised Languages (GLs).

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`
What evolved first: Languages for communicating, or languages for thinking
(Generalised Languages: GLs)

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0703`
Computational Cognitive Epigenetics

More papers/presentations are on my web site.

Can we give ACs suitable GLs ?

ACs without a similar pre-communicative form of representation allowing them to represent things about their world, and also possible goals to achieve, will not have anything to communicate when they start learning a language.

Their communicative competences are likely to remain shallow, brittle and dependent on pre-learnt patterns or rules for every task because they don't share our knowledge of the world we are talking about, thinking about, and acting in.

Perhaps, like humans (and some other altricial species), ACs can escape these limitations if they start with a partly 'genetically' determined collection of meta-competences that continually drive the acquisition of new competences building on previous knowledge and previous competences: a process that continues throughout life.

The biologically general mechanisms that enable humans to grow up in a very wide variety of environments, are part of what enable us to learn about, think about, and deal with novel situations throughout life.

I conjecture that this requires an architecture that grows itself over many years partly as a result of finding out both very general and very specific features of the environment through creative play, exploration and opportunistic learning.

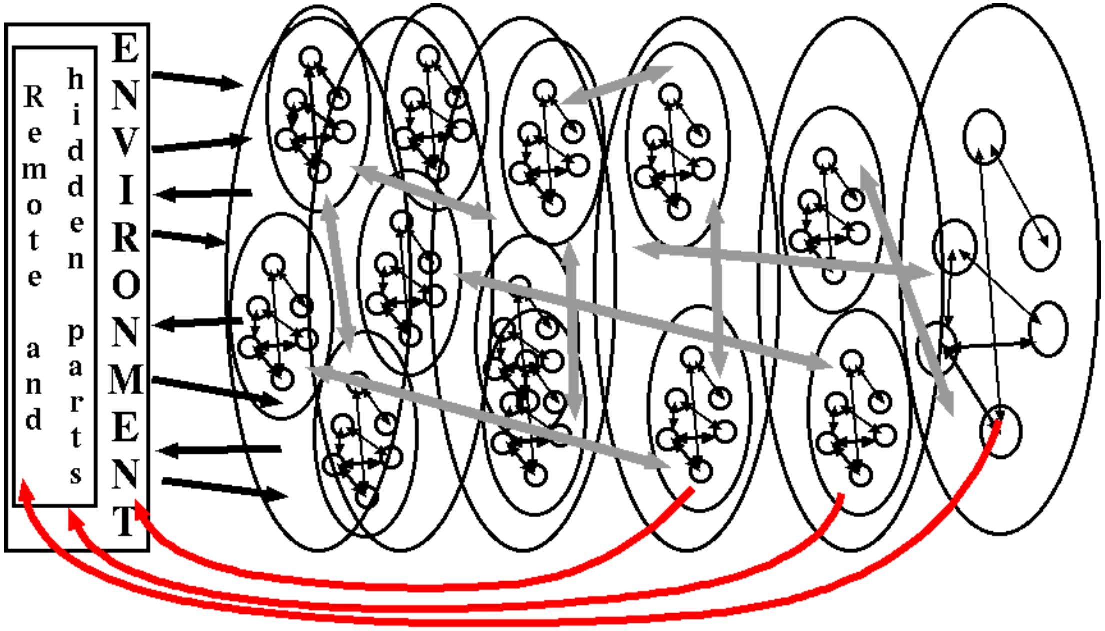
An architecture that builds itself

The hypothesised architecture is made of multiple dynamical systems with many multi-stable components, most of which are inactive at any time.

Some change continuously, others discretely, and any changes can propagate effects in parallel with many other changes in many directions.

Some components are closely coupled with environment through sensors and effectors, others much less coupled – even free-wheeling, and unconstrained by embodiment (and some of them can use logic, algebra, etc., when planning, imagining, designing, reminisicing, reasoning, inventing stories, etc.).

The red arrows represent theory-based semantic links from decoupled parts of the architecture to portions of the environment that are not necessarily accessible to the senses. See



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

Most animals have a genetically pre-configured architecture: the human one grows itself, partly in response to what it finds in the environment.

For more on this see

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0801

Architectural and representational requirements for seeing processes and affordances

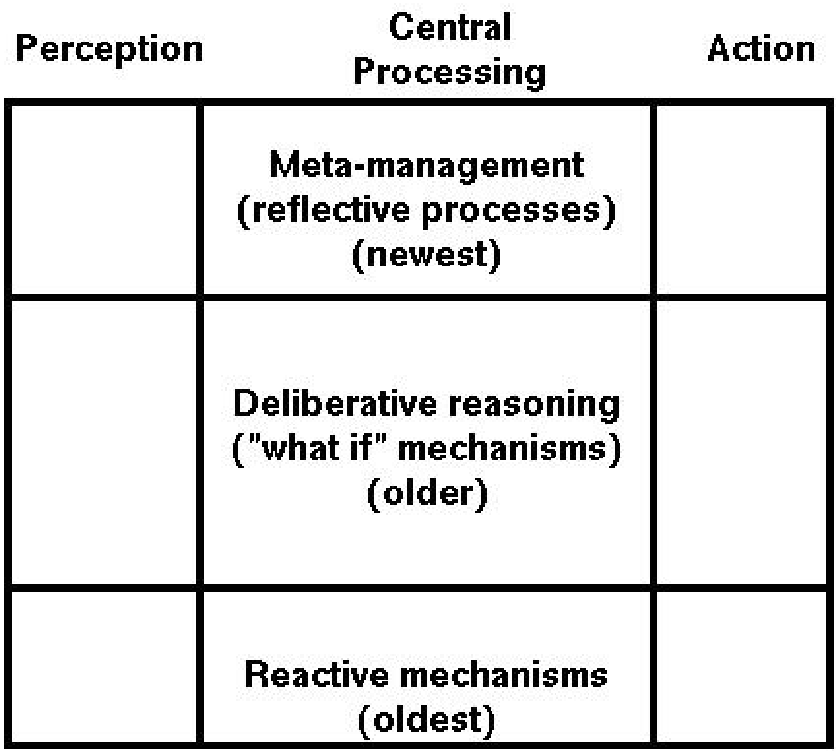
We can discern some major sub-divisions within a complex architecture

The CogAff Schema – for designs or requirements for architectures.

Different layers correspond to different evolutionary phases.

NEWER

OLDER



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

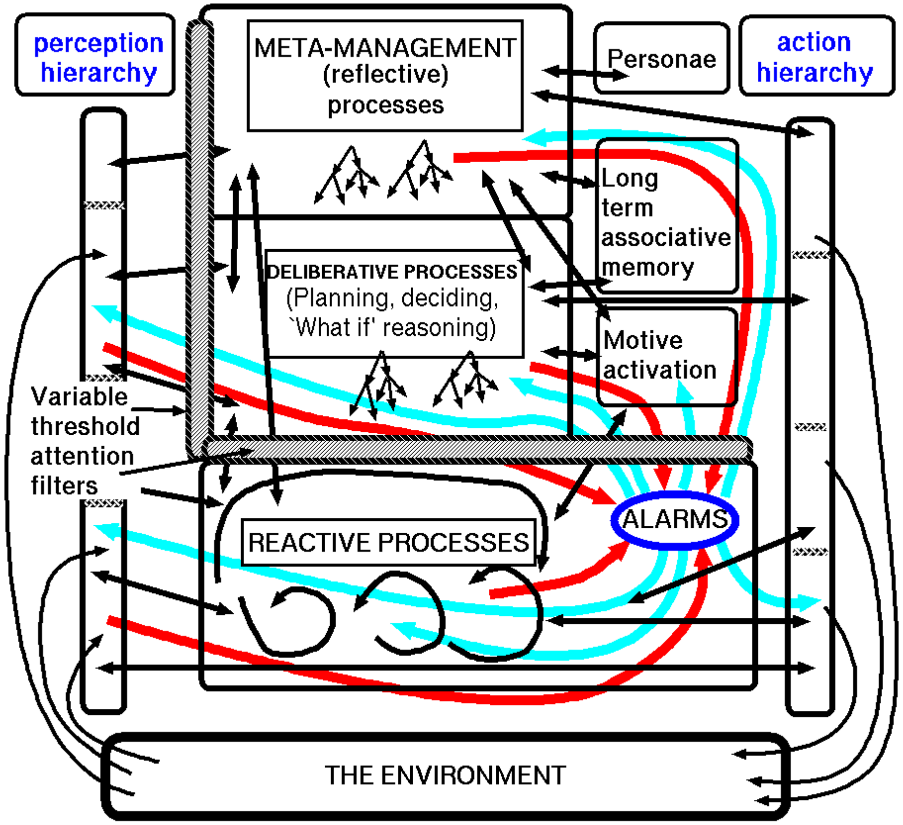
(Think of the previous “dynamical systems” diagram as rotated 90 degrees counter-clockwise.)

H-Cogaff: the human variety

Here’s another view of the architecture that builds itself (after rotating 90 degrees).

NEWER

OLDER


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

This shows a possible way of filling in the CogAff schema on previous slide – to produce a human-like architecture (H-CogAff). (Arrows represent flow of information and control)

For more details see papers and presentations in the Cogaff web site:

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

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks`

`http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200307` (progress report)

Machines that have their own motives

Human learning is based in part on what humans care about and vice versa: what humans care about is based in part on what they learn.

In particular, learning to help others depends on caring about what they want, what will and will not harm them, etc.

And learning about the needs and desires of others can change what we care about.

The factors on which the caring and learning are based can change over time – e.g. as individuals develop or lose capabilities, acquire new preferences and dislikes, etc.

If we make machines that can come to care, then that means they have their own desires, preferences, hopes, fears, etc.

In that case, we shall have a moral obligation to take account of their desires and preferences: anything with desires should have rights: treating them as slaves would be highly immoral.

As noted in the Epilogue to *The Computer Revolution in Philosophy* (1978), now online here:

`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`

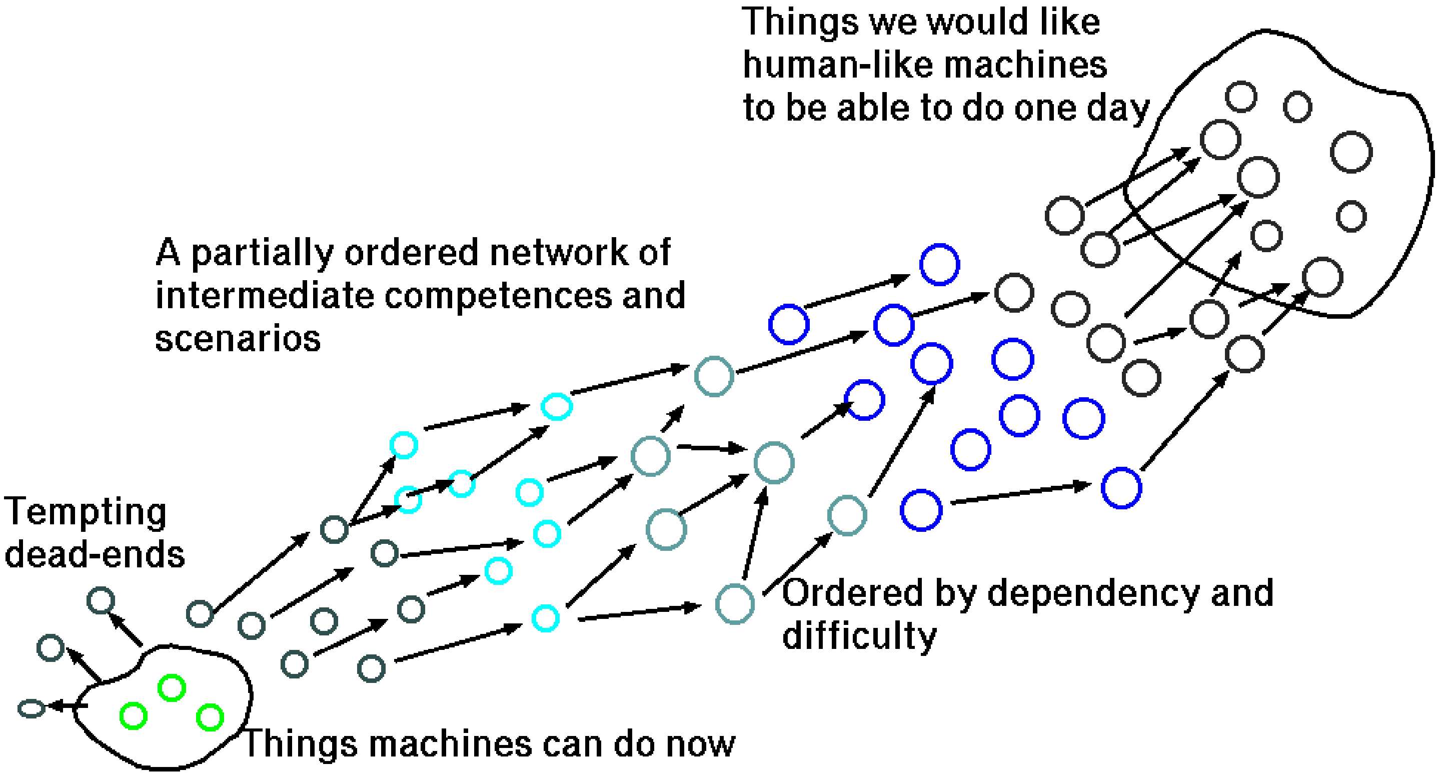
And in this paper on why Asimov's laws of robotics are immoral:

`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/asimov-three-laws.html`

Can it be done?

Perhaps we should work on this roadmap as a way of understanding requirements?

We can take care to avoid the stunted roadmap problem.



Things we would like
human-like machines
to be able to do one day

A partially ordered network of
intermediate competences and
scenarios

Tempting
dead-ends

Ordered by dependency and
difficulty

Things machines can do now

Don’t always go for small improvements – that’s a route to dead ends.
Try to envision far more ambitious goals, and work backwards.
Research planning should use more backward chaining.

Rights of intelligent machines

If providing effective companionship

requires intelligent machines to be able to develop their own

goals, values, preferences, attachments etc.,

including really *wanting* to help and please their owners,

then if some of them develop in ways we don't intend,

will they not have the right to have their desires considered,

in the same way as our children do

if they develop in ways their parents don't intend?

See also
`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/epilogue.html`

Recruiting needed?

When the time comes,

who will join the society

for the prevention of cruelty to robots?

When a young robot decides that it wants to study philosophy in order to be able to think clearly about how to look after people who need care, will Scottish universities let it in, without demanding a top up fee?

I hope so!

For more information see:

`http://www.cs.bham.ac.uk/research/projects/cogaff/`
The cognition and affect project web site

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks`
Online presentations (mostly PDF)

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/`
Papers produced in the CoSy robot project