

A (possibly?) New Theory of Vision
(and other modes of perception)
Combining Several Old Theories
Generating many new problems and research tasks.

Aaron Sloman

<http://www.cs.bham.ac.uk/~axs>
School of Computer Science, The University of Birmingham

With help from colleagues on the CoSy project

<http://www.cs.bham.ac.uk/research/projects/cosy/>
(Maria Staudte at DFKI kindly commented on an early draft)
And others, including Jackie Chappell, Biosciences, Birmingham.

These slides are accessible from here:

<http://www.cs.bham.ac.uk/research/cogaff/talks/>
<http://www.cs.bham.ac.uk/research/projects/cosy/papers/>
Along with other related slide presentations and papers.

New Theory Vision Slide 1 Last revised: February 17, 2007 Page 1

Acknowledgement

These slides make use of cartoons from a collection of cartoons entitled *French Cartoons* edited by William Cole and Douglas McKee, published in 1955 by Panther Books.

NOTE:

I have not found a way to contact the publisher to get permission to use these cartoons.

Perhaps my use of them for an academic purpose will be regarded as acceptable if it prompts new readers to look for the book.

New Theory Vision Slide 2 Last revised: February 17, 2007 Page 2

The (Possibly) New Theory

A HIGH LEVEL OVERVIEW OF THE THEORY

Vision is a process involving multiple concurrent simulations at different levels of abstraction in (partial) registration with one another and sometimes (when appropriate) in registration with visual sensory data and/or motor signals.

What all that means is explained more fully later.

The theory has different facets, which link up with many different phenomena of everyday life as well as experimental data, and with a host of problems in philosophy, psychology (including developmental and clinical psychology), neuroscience, biology and AI (including robotics).

If true, and possibly even if it is not true, it raises many new questions for all those disciplines and some others (e.g. linguistics).

Example: watch this video of child playing with a toy train set.

http://www.cs.bham.ac.uk/~axs/fig/josh34_0096.mpg

NOTE: This is work in progress.

Draft contents: likely to change

Contents

1 Preliminaries	6
2 Example: A child playing with a train-set on the floor	10
3 Background: precursors and the context of the CoSy project	13
4 Not just vision – compare speech and music	17
5 What we can see goes far beyond what we can do	21
6 The importance of concurrency	23
7 From structures (in the Popeye system) to processes	25
8 What if Popeye had been applied to a moving scene?	29
9 Cartoons and miming	33
10 Perceiving causation	38
11 Geometry-based causation	44
12 Multi-modal perception of causation	47
13 Many distinct competences have to be learnt	51
14 Much of what is learnt is about kinds of stuff	55
15 Viewpoint matters - some viewpoints are 'vicarious'	59
16 No good theories about shape perception exist	65
17 Not only humans	69
18 A child can appear less competent than a crow	72
19 Running 2-D or 3-D simulations to answer questions	74
20 Buggy simulation in a five year old child	76
21 Seeing structure, motion, and invariants in mathematics	80
22 High level perceptual processes can ignore low-level details	83

23	What the simulation theory does and does not say	87
24	Seeing structures, processes and causation	91
25	What I am NOT saying	93
26	Some inadequacies of closely related theories	98
27	First steps towards clarifying terminology	100
28	Re-runnable check-points	102
29	How the theory arose in the context of the CoSy project	104
30	Seeing and acting on everyday objects like cups, spoons and saucers	109
31	Orthogonal competences (Added 1 Jan 2006)	114
32	Potential applications of a child-like robot with abilities described here	123
33	We need to think about architectures	125

1 Preliminaries

First some preliminary remarks about the scope of the ideas.

A Shift of View

- For many years I assumed (like many other people) that if we could understand perception of static scenes we could later deal with motion.
- I also thought (as explained below) that perception of a static scene involved forming (static) descriptions of its contents (at different levels of abstraction), and that a theory of perception of motion might later be derived from that.

A Shift of View

- For many years I assumed (like many other people) that if we could understand perception of static scenes we could later deal with motion.
- I also thought (as explained below) that perception of a static scene involved forming (static) descriptions of its contents (at different levels of abstraction), and that a theory of perception of motion might later be derived from that.
- Then I learnt about Gibson's theory of affordances, which made it necessary to relate perception of static scenes to the *possibility of* (and constraints on) actions and their consequences that are not occurring, but might occur.
- For a while I assumed that a theory of perception of affordances could be tacked onto a theory of perception of structure by representing the perceived affordances as collections of something like condition-action rules associated with various parts of a scene.

A Shift of View

- For many years I assumed (like many other people) that if we could understand perception of static scenes we could later deal with motion.
- I also thought (as explained below) that perception of a static scene involved forming (static) descriptions of its contents (at different levels of abstraction), and that a theory of perception of motion might later be derived from that.
- Then I learnt about Gibson's theory of affordances, which made it necessary to relate perception of static scenes to the *possibility of* (and constraints on) actions and their consequences that are not occurring, but might occur.
- For a while I assumed that a theory of perception of affordances could be tacked onto a theory of perception of structure by representing the perceived affordances as collections of something like condition-action rules associated with various parts of a scene.
- In retrospect it seems silly to have forgotten that vision evolved in organisms embedded in a dynamically changing environment – so its primary function must be not to discover **what exists** in the environment, but **what is happening** in the environment, including the perceiver's movements and actions.
- Add the observation that what is happening, and what is potentially important to an organism, is not a unique process, but a collection of processes at different levels of abstraction, e.g. a wave moving horizontally towards the shore and millions of molecules mostly moving roughly up and down in the same place.

2 Example: A child playing with a train-set on the floor

The video mentioned above shows a child about three and a half years old doing things with a train set that surrounds him as he sits in the middle, turning this way and that, pointing at things behind him to answer questions, pushing the train through a tunnel, changing his position to replace a tree knocked down by the back of his head when he puts his head down to look through the tunnel.

http://www.cs.bham.ac.uk/~axs/fig/josh_tunnel.mpg (5MB)

http://www.cs.bham.ac.uk/~axs/fig/josh_tunnel_big.mpg (15MB)

(High resolution version.)

My claim that this child is running various simulations of things going on in the environment begs the question: 'What kind of thing is a simulation?'

My provisional answer is that anything that is capable of usefully representing a process can be called a simulation for present purposes, even if it is a static structure accessed sequentially.

Later I'll say more about what I do and do not mean.

NOTE: I am not making any use of Grush's distinction between 'emulation' and 'simulation', though it is possible that it will turn out that what I mean by 'simulation' is what he means by 'emulation'.

Snapshots from tunnel video

A child playing with his train illustrates the theory.



- The child clearly knows what's going on in places he cannot see.
- He can point at and talk about something behind him that he cannot see.
- When he turns to continue playing with the train he knows which way to turn and roughly what to expect.
- When the train goes into the tunnel and part of it becomes invisible, he does not see the train as being truncated, and he expects the invisible bit to become visible as he goes on pushing.
- He sees the whole train as one thing while part of it is hidden in the tunnel.
- What is the role of **vision** in all of this? Frequently sampling the environment?

Not all of this competence is there from birth: at least some of it has to be learnt: what does that involve and what mechanisms make it happen?

New Theory Vision

Slide 8

Last revised: February 17, 2007 Page 11

Tunnel vision

Think about the child playing with and talking about his toy train, with track, tunnel and other things on the floor around him.

How many different levels of abstraction occur in

- the processes he needs to perceive,
- the processes he needs to use in controlling his actions,
- the processes he needs to think about, explain, modify, predict, ...

Is there a sharp division between

- seeing geometric structures, relationships, changes and
- seeing causal and functional relations?

Is there a sharp distinction between what the child sees as **caused by his action**, and what he sees as merely **happening in the environment**?

Could the same mechanisms represent both?

Compare

- Movement of the truck he is holding and pushing
- Movement of the truck adjacent to the one he is pushing
- Movement of the trucks in the tunnel that cannot be seen
- Reappearance of the front of the train from the far end of the tunnel

We return to perception of causal relations later.

New Theory Vision

Slide 9

Last revised: February 17, 2007 Page 12

3 Background: precursors and the context of the CoSy project

Many previous observations and theories lie behind the ideas being presented here.

The trigger that led me to think this way about perception, combining a lot of old ideas in new way, was working on the EC-funded CoSy project, in particular the 'PlayMate' robot scenario in which a robot manipulates 3-D objects on a table top:

<http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html>

As I started to analyse in great detail what needs to be represented when a robot moves one complex structured object in relation another complex structured object (rigid or flexible), I realised that complex objects involve 'multi-strand' relationships, and when they move that involves 'multi-strand' [processes](#).

Background

- There are many views of the nature and function(s) of vision, including the following:
 - Vision produces information about physical objects and their geometric and physical properties, relationships in the environment.
(Marr and many others.)
 - Much recent work treats vision as a combination of recognition, classification and prediction – the latter sometimes used in tracking
(often using classifications arbitrarily provided by a teacher, rather than being derived from the perceiver's needs and the environment).
 - Vision controls behaviour (Obviously true?)
 - Behaviour controls perception, including vision. (W.T.Powers)
 - Vision is unconscious inference (Helmholtz)
 - Vision is controlled hallucination (Max Clowes) [Pretty close](#)
- I'll try to present phenomena that require a richer deeper theory.
 - It will be evident that the new theory uses many of the above ideas, and assembles them with some new details. Some of the ideas are criticised.
 - The implications seem to be very important: both for studies of vision and cognition in animals (especially, but not only humans), and for attempts to understand requirements for robots with human-like capabilities.

Relationship with CoSy project

A change of view came while I was working on the CoSy project

<http://www.cs.bham.ac.uk/research/projects/cosy/>

I have been thinking about many of the problems for many years, but what made things click into place recently was examining very closely the perceptual and representational requirements for a robot manipulating 3-D objects on a table-top, e.g. watching a hand picking up a cup, or assembling a meccano model.

Try thinking about it yourself!

Using one or two hands, perform simple, everyday actions on cups, spoons, scissors, paper, string, a handkerchief, nuts and bolts, tin-openers, your food, a sweater you put on or remove

and watch very, very closely.

How can your brain represent the information you use, including

- all the things and processes you see, as complex 3-D objects move while changing their shapes and mutual relationships,
- what you anticipate,
- your recollection of what just happened,
- your thoughts about what would have happened if you, or someone else, had done something different?

PERHAPS YOU WILL INVENT THE SAME THEORY.

The theory is not totally new

There are many precursors of different kinds:

Some old philosophical theories of minds as idea-manipulators.

Kant's *Critique of Pure Reason* (1780) (Including his theory of mathematical knowledge)

Helmholtz: perception is unconscious inference

Kenneth Craik in 1943 (animals use predictive models)

Ulric Neisser and others (1960s): theories of vision as analysis by synthesis, and hierarchical synthesis.

Karl Popper (our hypotheses can die in our stead)

William T Powers: Behaviour controls perception.

Lots of control engineering using 'predictive' models.

Max Clowes: Vision is controlled hallucination

David Hogg's work on perceiving a walking person (1983)

My own work in the 1970s on multi-level perception and visual reasoning

Work by Tsotsos on motion perception.

Roger Shepard and others on mental rotation tasks.

Steve Kosslyn on imagery

Phil Johnson-Laird on reasoning with mental models

JJ Gibson on perceiving affordances (and his earlier ideas about 'perceptual systems')

Minsky's *Society of Mind* and other work.

Arnold Trehub: (1991) *The Cognitive Brain*

Alain Berthoz (2000) *The Brain's sense of movement*,

Murray Shanahan AISB 2005

[R. Grush, 2004, The emulation theory of representation: ... BBS, 27,](#)

And probably more: but does any combine all the elements proposed here?

(Grush comes closest)

4 Not just vision – compare speech and music

I was thinking mainly about vision when I started writing these ideas down.

I soon realised that the points could be generalised to other forms of perception and to multi-modal perception, e.g. seeing feeling and hearing the same thing.
(Which would typically be represented a-modally.)

So the presentation is really about aspects of cognition in humans, human-like robots and possibly some other animals.

But I have left the original title.

So, although this talk is mainly about vision, not about other forms of perception, there seems to be a lot in common between the role of simulation capabilities in the ability to experience things, imagine things, invent things, reason about things, and there are commonalities between vision and other perceptual modalities, e.g. haptic perception, auditory perception, auditory imagination, auditory composition.

Moreover, like Grush, I'll give amodal and multi-modal examples.

Not just vision – consider music

- What happens when you hear music?
- Some music is in principle something you could produce yourself, e.g. a person singing.
- So hearing singing might involve a process that activates internal parts of your singing capability.
- But what if it's a duet, or a trio, or a singer accompanied by orchestra, or a string quartet?
 - You can hear and enjoy music without activation of incipient actions you could produce: many who enjoy listening to instrumental music have never learnt to play an instrument.
 - Some musicians can also judge scores without hearing the music played: they are not judging pretty patterns on the paper – probably something like a simulated performance.
 - Some musicians can play, and listen, in a group, all improvising concurrently. How?
- So we have an ability to experience and appreciate processes that are richer and more complex than anything we can produce using our own bodies.

Not just vision – consider music

- What happens when you hear music?
- Some music is in principle something you could produce yourself, e.g. a person singing.
- So hearing singing might involve a process that activates internal parts of your singing capability.
- But what if it's a duet, or a trio, or a singer accompanied by orchestra, or a string quartet?
 - You can hear and enjoy music without activation of incipient actions you could produce: many who enjoy listening to instrumental music have never learnt to play an instrument.
 - Some musicians can also judge scores without hearing the music played: they are not judging pretty patterns on the paper – probably something like a simulated performance.
 - Some musicians can play, and listen, in a group, all improvising concurrently. How?
- So we have an ability to experience and appreciate processes that are richer and more complex than anything we can produce using our own bodies.
- This implies that if perception is simulation, there must be some simulation mechanisms that are **not very closely tied to details of action mechanisms**.

Not just vision – consider language

- Humans can talk, and read aloud, but they can also learn to read silently. It is commonplace to assume that reading silently involves somehow suppressing the final stages of the process of reading aloud.
- But perhaps we'll discover that it is better to construe reading silently as a process of **simulation** of at least three sorts of things:
 - (a) of reading aloud
 - (b) of listening to someone else talk
 - (c) of the situations, actions, events described or narrated.
- When we hear other people talk we (mostly unconsciously) analyse and interpret the sounds they make, building interpretations of different sorts concurrently.
 - (e.g. in understanding a story — information is processed at different levels of abstraction including phonemes, words, phrases, sentences, story-plots, so we hear, a stream of sounds, a stream of words, streams of higher level syntactic structures, streams of ideas, speech acts, events described...)

5 What we can see goes far beyond what we can do

As shown by previous examples we can see and think about things we cannot do or construct

(e.g. when we reason about transformations of infinite sets, like reversing all the even numbers and placing them after the odd numbers).

So any theory of human mental processing that tries to relate all mental processes to relations between perception and action must be wrong.

Simulation capability exceeds behavioural capability

If human brains (and perhaps others) can construct and run simulations of processes of many kinds, there is no need for each one to be **closely** related **either** to the specific motor system that would be used to produce such processes **or** to the sensory systems that would be used to perceive such a process.

After all, we can perceive many processes we cannot produce, e.g. waterfalls – and we shall later give examples of perceiving and thinking about ‘vicarious affordances’, i.e. affordances for others.

So we have an ability to experience and appreciate processes that are richer and more complex than anything we can produce using our own bodies. As stated above: if perception is simulation, there must be some simulation mechanisms that are **not very closely tied to details of action mechanisms**.

- Evolution apparently ‘discovered’ the benefits of structural and causal disconnection between representation and thing represented, long ago (in a subset of animals only?): can we replicate this in our designs?
- Compare
 - the ability of a prey animal to think about what a predator might do
 - the ability of a composer to think up a multi-performer composition, and specify it in a musical score.
 - the ability of a general to prepare orders for various concurrently active platoons.
 - the ability of some programmers to design, implement, and debug programs involving concurrent processes (e.g. operating systems).

6 The importance of concurrency

What we can see (and what we can represent) can involve multiple concurrent processes including processes at different levels of abstraction.

The importance of concurrency

Besides emphasising the importance of **processes** as being the content of what is perceived (i.e. not just static structures), we are also emphasising the importance of **concurrency**, namely the perception as involving multiple perceived processes, some at the same level of abstraction, some at different levels of abstraction

- Perceived concurrency is involved in various human and animal activities involving two or more individuals engaged in fighting, dancing, mating, playing games, performing music, etc.
- Doing this well implies a need to be able to keep track of (partly by running simulations?) the actions of others at the same time as planning and performing one's own actions.
- What are the evolutionary precursors of this, e.g. in hunting animals and prey of hunting animals, including parents defending young from predators?
- Concurrency is also important in social learning, since many social interactions are concurrent rather than simply based on turn-taking: e.g. dancing, old friends embracing, lifting or pushing a heavy article, and mating.
- **Conjecture:** our architecture evolved to support at least three sorts of concurrency:
 - Perceiving multiple concurrent external processes
 - Representing the same process at different levels of abstraction
 - Different concurrent actions in an individual, such as walking (including posture control), working out where to walk, discussing philosophy with a companion, using different parts of the information-processing architecture.

7 From structures (in the Popeye system) to processes

About 30 years ago a project at Sussex University explored some aspects of the theory that perception of complex and noisy structures could be facilitated by a visual architecture in which processes at different levels of abstraction, concerned with different ontologies, ran concurrently with a mixture of bottom up and top down control, including top-down control of attention.

But there was nothing in this about perceiving **processes** at different levels of abstraction, as is proposed here. Yet some of the ideas remain relevant.

An example old idea that's still relevant

Around 30 years ago I was working with David Owen and Geoffrey Hinton on a theory of vision that involved multi-level interpretation of static images, as on the next slide.

The theory explained how high level decisions could be reached relatively robustly and quickly, despite considerable complexity and noise at lower levels.

- If high level decisions were derived directly from low level image details the search space would be astronomical.
- By finding intermediate level recognisable structures and using their relationships to trigger high level hypotheses, while higher levels controlled 'attention' and some thresholds at lower levels, we allowed the sparsity of high level models to drive both speed and robustness. (U. Neisser called this 'Analysis by synthesis' about 40 years ago. Later it was called 'heirarchical synthesis'. It has probably been reinvented many times.)
- The system degraded gracefully in both speed and accuracy as noise and clutter were added at the lowest level.
- A working implementation of that idea, called 'POPEYE' was described in chapter 9 of '*The Computer Revolution in Philosophy*' (<http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html>)
- On that view, seeing involved creating multi-level **structures** concurrently.

The next slide illustrates this old idea, showing how the Popeye program interpreted pictures made from dots by analysing the picture at different levels of abstraction in parallel, each level involving a different ontology from the others, using a mixture of bottom-up (data-driven) and top-down (model-driven, hypothesis-driven) interpretation, with rich structural relations between details at different levels.

Multiple levels of structure perceived in parallel

Old conjecture: We process different layers of interpretation in parallel.

Obvious for language. What about vision?

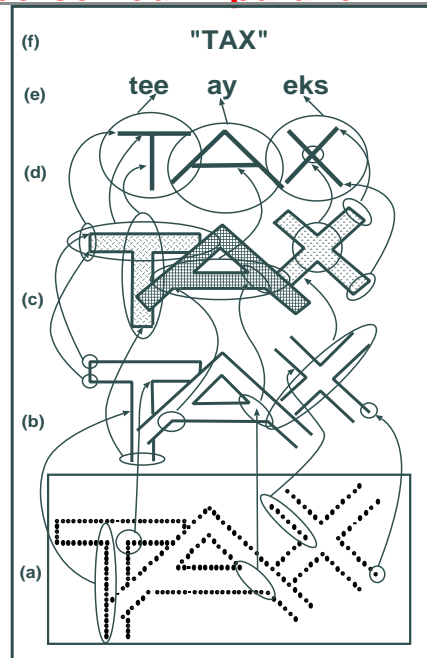
Concurrently processing bottom-up and top-down helps constrain search. There are several ontologies involved, with different classes of structures, and mappings between them – so the different levels are in 'partial registration'.

- At the lowest level the ontology may include dots, dot clusters, relations between dots, relations between clusters. All larger structures are **agglomerations** of simpler structures.
- Higher levels are more abstract – besides **grouping** (agglomeration) there is also **interpretation**, i.e. mapping to a new ontology.
- Concurrent perception at different levels can constrain search dramatically (POPEYE 1978) (This could use a collection of neural nets.)
- Reading text would involve even more layers of abstraction: mapping to morphology, syntax, semantics, world knowledge

From *The Computer Revolution in Philosophy* (1978)

<http://www.cs.bham.ac.uk/research/cogaff/crp/chap9.html>

Replace all that with concurrent multi-level processes – using different process-ontologies.



From Structures to Processes

I now propose to replace the idea that

1. seeing involves multi-level structures in partial registration using different ontologies, with the claim that
2. seeing involves multi-level process-simulations in partial registration using different ontologies, with rich (but changing) structural relations between levels.

- Shortly after the work on Popeye was done, David Hogg was a PhD student in the same department working on motion perception.

D. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

- His well known 'walking man' system was an early example of what I am now talking about: his model-based interpretation of a video of a walking man amounted to a simulation of a walker, partly controlled by the changing image data, and partly controlled by the dynamics of the model.
- Despite being his supervisor I did not appreciate the full significance of that work till now.

I think he also did not see the full significance of what he had done: he described the system as showing how to use a model to interpret an image, rather than claiming to show how to interpret a sequence of images as representing a process.

8 What if Popeye had been applied to a moving scene?

At the time Popeye was developed we were not thinking about motion perception (although other people were, partly inspired by J.J.Gibson's ideas about the importance of optical flow).

It is interesting to reflect on what we might have done differently had we used pictures of moving (e.g. sliding rotating) laminas with similar kinds of overlap and noise.

What we did not do in the Popeye program

- We did not develop a program capable of representing the same multi-level structures, but with the objects in constant motion.
- An experiment to try one day would be producing movies derived from the 'dotty' word representing pictures. The conjecture is that people would not only see moving dots, but also moving lines, moving laminas, moving letters, thought it is not exactly clear how this would be objectively tested.
- I suspect we could cope with relative motions of parts of letters, e.g. so that the angles between parts of the letters change.

Compare the work of Gunnar Johansson on movies made by attaching lights to joints on people, and filming them moving in the dark: when the lights start moving a 3-D process is perceived.

Excellent demo: <http://www.bml.psy.ruhr-uni-bochum.de/Demos/BMLwalker.html>

Changes required for switching the Popeye architecture to a moving scene

- It would be silly to keep all the low level detail indefinitely as new details would be coming in all the time
- Different times of preservation would be relevant to different things at different levels in the ontology, e.g. depending on whether they are large or small, static or moving, or of interest relative to some goal.
- It might be useful to add low level motion maps, or even to replace the static low level maps completely.

(Compare A.Trehub: *The Cognitive Brain*)

How to see a static scene as a process

If all this is right, our ability to see processes is used even when we look at a static scene:

it's just that then the process is one in which nothing changes.

- But if something started changing we would see it, using the same mechanisms as were previously perceiving the static configuration.
- A static scene is just a special kind of process, in which nothing changes.
- Whether the things change or not the system has to be prepared for many possibilities.
- Thus perception of a static structure already involves perception of possibilities for motion (mostly latent: the simulation capabilities may be turned on if motion occurs, and left dormant otherwise).

This could be seen as a minimal notion of affordance.

Reminding the audience: relevant things you probably know

There are many aspects of our everyday experience that people may or may not notice that seem to involve this ability to run some sort of simulation of environmental processes.

So this is not really a theory that's new to you, even if you previously never thought about it.

- E.g. when you see something moving behind an opaque object you don't see the moving object as being truncated – you see it as having a hidden portion that continues to move (like the child in the video pushing his train into a tunnel), and typically you know roughly where the hidden parts are as the motion continues (though of course stage conjurers can fool us because we are not infallible).
- Many cartoons and jokes depend on our ability to run simulations derived from the information presented, e.g. pictorially or verbally.
- Doodles depend on this ability too. In fact many/most(?) forms of visual art do.

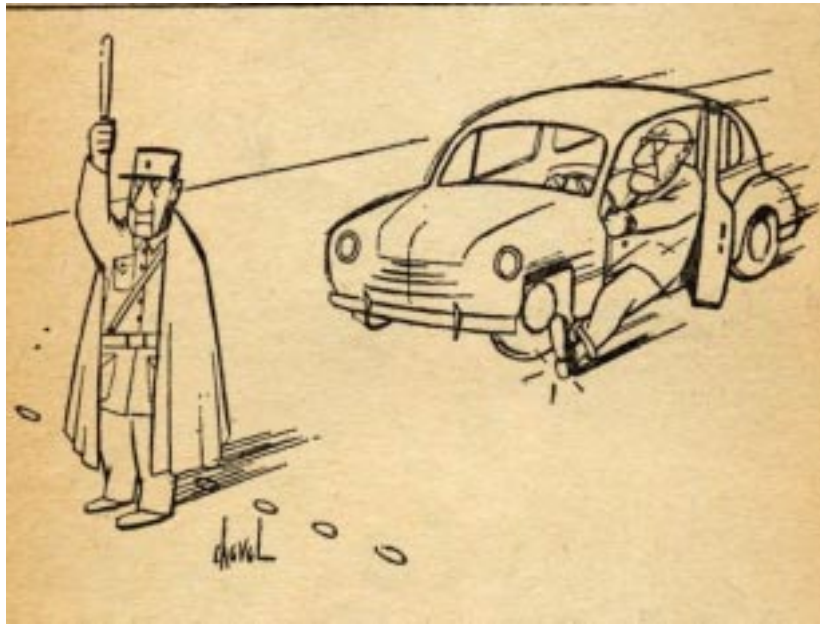
Some cartoons showing 'snapshots' of extended processes follow. Some project into both future and past, some only one or the other.

'French Cartoons' Published 1955

Ed William Cole and Douglas McKee, : Panther Books

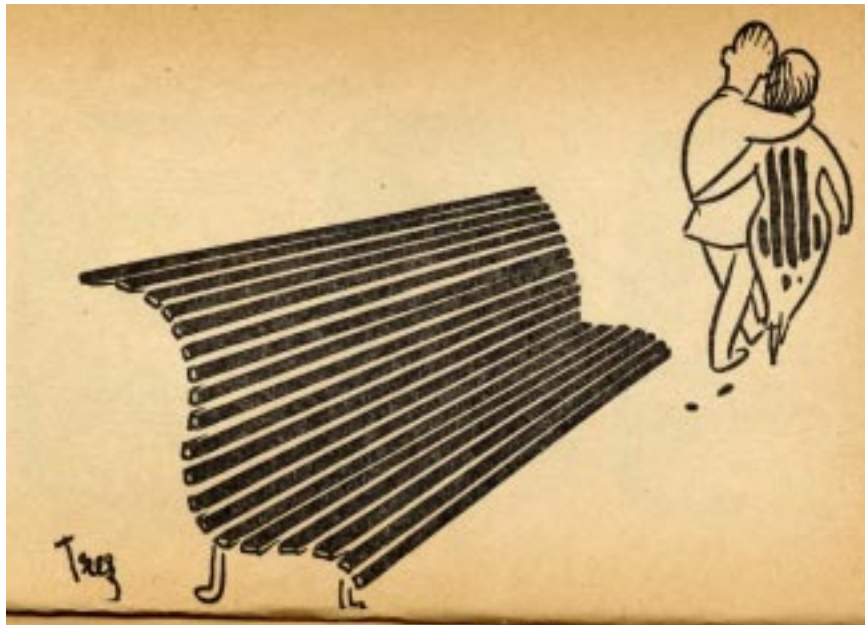
When you look at the cartoons that follow, what past and future processes come to mind and how do they relate to details of the scene?

Mostly future



Another kind of footbrake ???

Mostly past



Understanding the picture involves 'running a simulation' but at a high level of abstraction with many details of the previous history left out.

We produce external simulations also

Using a tennis ball and badminton shuttlecock to simulate eating an ice-cream – he never actually licked the ball.

We often use external simulations, including gestures, diagrams, working models. However most of our examples below will be cases of purely internal simulation.

Perhaps a major function of play in young mammals is developing simulation capabilities through learning about different things to simulate (as opposed to developing motor skills, muscles, etc.)



Evolution (and processes in individual development) somehow gave us the ability to make use of either **internal** or **external** objects, when running simulations. My 1971 IJCAI paper claimed that reasoning with diagrams is essentially the same thing whether done **on paper** or **in the mind**.

Brain mechanisms for this are still waiting to be discovered.

(See the interesting discussions in BBS paper and commentary by R.Grush, 2004 – found after much of this had been written).

10 Perceiving causation

Two kinds of causation: Humean (probabilistic, evidence based) and Kantian (deterministic: based on hypothesised structures)

Perceiving causation

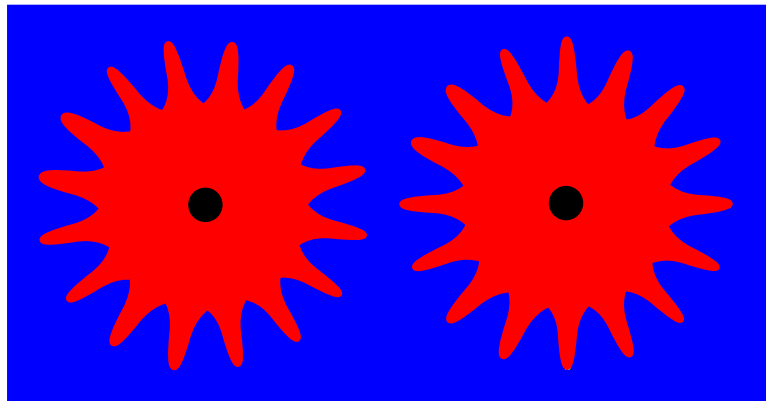
Our ability to perceive moving structures, and our meta-level ability to think about what we perceive, is intimately bound up with perception of causation and affordances.

In some cases the causal relations are inherent in what is seen, whereas in others they involve invisible structures and processes: but the same key idea is used in both cases.

Illustrations follow.

Invisible, Humean, causation – mere correlation

Two gear wheels attached to a box with hidden contents.
Here we do not perceive causation.



Can you tell by looking what will happen to one wheel if you rotate the other about its central axis?

You can tell by experimenting: you may or may not discover a correlation.

Compare experiments reported by Alison Gopnik in her invited talk at IJCAI'05, Edinburgh July 2005

Visible, intelligible, Kantian, causation

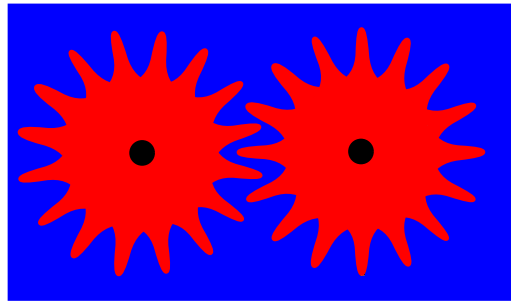
Two more gear wheels:

Here you (and some children) can tell 'by looking' how rotation of one wheel will affect the other.

NB The simulation that you do makes use of not just perceived shape, but also **unperceived constraints**:

rigidity and impenetrability. These constraints need to be part of the

perceiver's ontology and integrated into the simulations, for the simulation to be deterministic.



Visible structure does not determine all the constraints: we also have to learn about the nature of materials, to see what is happening, and understand causation.

We need to explain how brains and computers can set up and run simulations involving multiple concurrent changes of relationships, subject to varying constraints determined by context.

These ideas are developed in two online documents

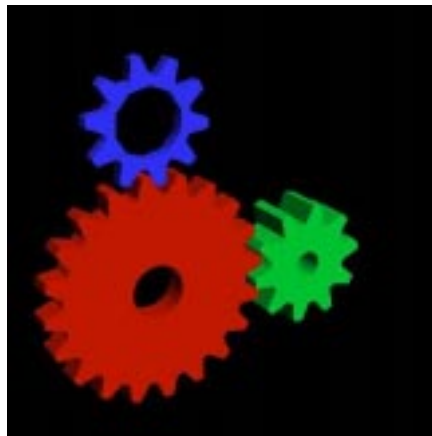
<http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0506>

COSY-PR-0506: Two views of child as scientist: Humean and Kantian

<http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0601>

COSY-DP-0601 Orthogonal Competences Acquired by Altricial Species (Blanket, string and plywood).

MORE GEARS



SHOW THE GLXGEARS DEMO

This is available on many linux systems.

If the central holes have fixed axles through them and the blue wheel turns clockwise, what will the others do?

What changes if one of the wheels slides along its axle while the others do not?

Humean and Kantian Causation

- When the only way you can find out what the consequence of an action will be is by trying it out to see what happens, you may acquire knowledge of causation based only on observed correlations. This is 'Humean causation' – David Hume said there was nothing more to causation than constant conjunction, and this is now a popular view of causation: causation as statistical (often represented in Bayesian nets).
- However if you don't need to find out by trying because you can see the structural relations (e.g. by running a simulation that has appropriate constraints built into it) then you are using a different notion of causation: Kantian causation, which is deterministic and structure-based.
- I claim that as children learn to understand more and more of the world well enough to run deterministic simulations they learn more and more of the Kantian causal structure of the environment.
- Typically in science causation starts off being Humean until we acquire a deep (often mathematical) theory of what is going on: then we use a Kantian concept of causation.
- **This requires learning to build simulations with appropriate constraints.**

For more on this see this talk

<http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0505>

COSY-PR-0506: Two views of child as scientist: Humean and Kantian

11 Geometry-based causation

Perceiving causation in changing geometric structures.

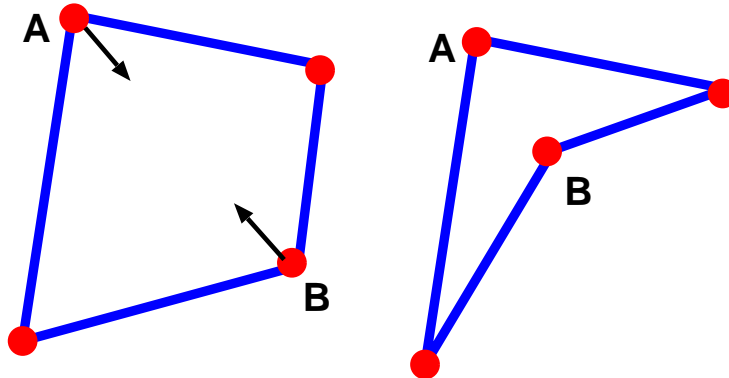
We can often see and understand consequences of motion of one part of a structure, including being able to predict effects on other parts.

But not when the structures are too complex, or have too many degrees of freedom.

Every kind of human competence has fairly low complexity limits, even though humans are enormously flexible in deploying and combining their competences.

Simulating motion of rigid, flexibly jointed, rods

On the left: what happens if joints A and B move together as indicated by the arrows, while everything moves in the same plane? Will the other two joints move together, move apart, stay where they are. ???



- What happens if one of the moved joints crosses the line joining the other two joints?
- We can change the constraints in our simulations: what can happen if the joints and rods are not constrained to remain in the original plane?

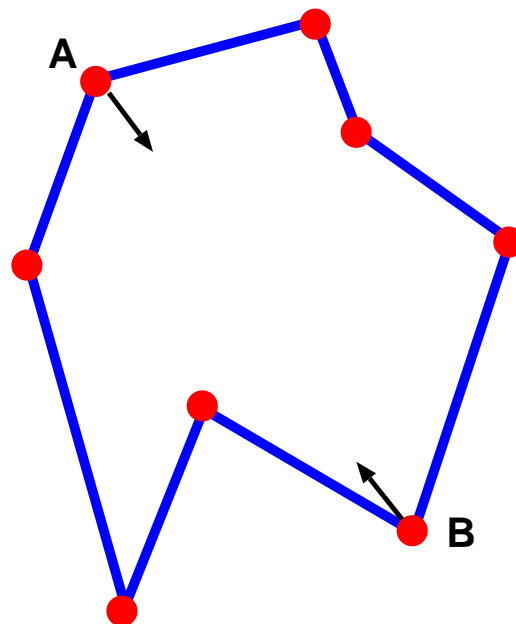
Multiple links: how we break down

Can you tell how the other rods will move, if A and B are moved together and all the rods are rigid, but flexibly jointed?

There are not enough constraints. In this case our causal reasoning merely allows us to think about a range of options, though it is not easy. Unlike simpler linkages, most people will not be able to see whether the continuum of possible processes divides into clearly distinct subsets except (perhaps) by spending a lot of time exploring.

As situations get more complex, human abilities to simulate degrade rapidly: our understanding of Kantian causation tends to be limited to relatively simple, deterministic cases, though we can learn to grasp more complex structures and processes – up to a point.

Perhaps intelligent artificial systems will have similar limitations.



12 Multi-modal perception of causation

We can combine information from different senses to produce a running simulation of what is going on.

(As Grush (2004) points out.)

In some case what is represented in the simulation is not sensed at all, until some time after the simulation starts.

Mixed mode input to an integrated simulation

- What you hear, like what you see, can be a process occurring in the environment, for instance hearing someone moving round you when your eyes are shut.
- If you are sitting in a room with a door opening into a corridor, subtle aspects of the changing sound of footsteps (which you process unconsciously) may produce a percept of an unseen person moving to the door, so that you know when he will become visible – a device used often in movies.
- Likewise when you see the unseen person's shadow changing.
- So the process you **hear** occurring and the things you **see** occurring may exist in the same integrated simulation — which is just as well since they exist in the same spatial environment.
- Likewise what a dentist sees and feels with the probe as she looks into the patient's mouth need to be in the same perceived part of the world, and when you use a hand to feel the underside of the table you are looking at **you see and feel the same table**.
- If you push a pencil up through a hole in the table you see and feel the same moving pencil.

Sensory modality and mode of representation

- Sensory modality driving a simulation need not determine the nature of the percept.
- A **unitary** percept of a process can be driven by input from **diverse** sensory modalities – e.g. seeing, hearing, feeling the same thing happening.
- What is simulated does not determine the nature of the medium used to implement the simulation, as long as it has a rich enough structure and appropriate mechanisms to create, modify, access and use the contents.
- Examples of what the simulation might be include:
 - a set of variables with changing values driven by sensory data
 - a database of logical assertions along with insertions and deletions driven by sensory data
 - a hybrid mechanism – logical assertions with equations linking changing variables, as can happen in some spreadsheets,
 - a spatially structured changing model,
 - a stored 'script' for the process with a pointer moving through the script at a rate determined by sensory input,
 - it may use a powerful form of representation that we have not yet thought of though evolution discovered it long ago.
- **Whatever form of representation is used, currently known brain mechanisms do not seem to support the required functionality.**

Visual reasoning about something unseen

An example of disconnection between simulation and sensory data.

If you turn the plastic shampoo container upside down to get shampoo out, why is it often better to wait before you squeeze?

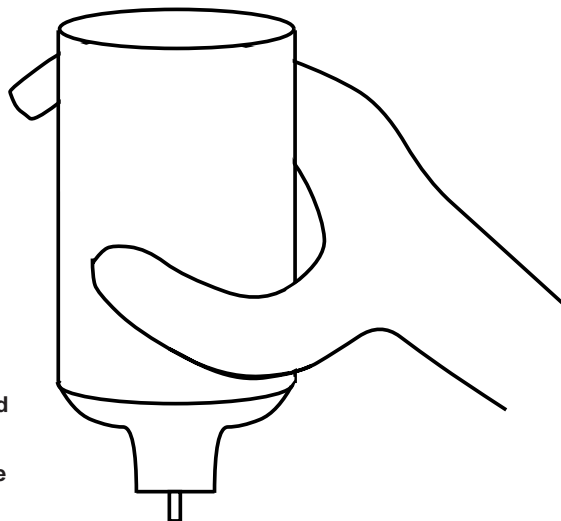
In causal reasoning we often use runnable models that go beyond the sensory information: part of what is simulated cannot be seen – a Kantian causal learner will constantly seek such models, as opposed to Humean (statistical) causal learners, who merely seek correlations.

Note that the model used here assumes uncompressibility rather than rigidity.

Also, our ability to simulate what is going on explains why as more of the shampoo is used up you have to wait longer before squeezing.

Sometimes we run the wrong simulation if we don't understand what is going on.

Like the person who suggested that you have to wait for the water from the shower to warm the air in the container.



13 Many distinct competences have to be learnt

The competences described above are not all present at birth, though some of the mechanisms required to acquire them are (while other learning mechanisms have to be produced by learning).

They are not **pre-configured** by genetic mechanisms, like innate abilities or innate latent genetically-determined competences that emerge long after birth (e.g. sexual competences, or migration in some birds).

The learnt, meta-configured competences need powerful bootstrapping mechanisms.

See

A. Sloman and J. Chappell (2005), The Altricial-Precocial Spectrum for Robots, *Proceedings IJCAI'05* pp. 1187–1192.

<http://www.cs.bham.ac.uk/research/cogaff/05.html#200502>

A. Sloman and J. Chappell (2005), Altricial self-organising information-processing systems, *AISB Quarterly*, 121, Summer 2005, pp. 5–7,

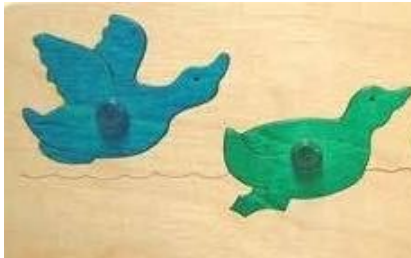
<http://www.cs.bham.ac.uk/research/cogaff/05.html#200503>

What the bootstrapping mechanisms achieve is extremely dependent on what is in the environment (including the culture), which is why altricial species with many meta-configured competences can differ enormously in what they know and can do, unlike precocial species, in which most competences are pre-configured, like deer which run with the herd soon after birth.

The examples that follow indicate some of what a child has to learn to see, before it can control its actions so as to achieve its goals, like inserting a puzzle piece where it belongs.

We cannot do it all from birth

The causal reasoning we find so easy is difficult for infants.



A child learns that it can lift a piece out of its recess, and generates a goal to put it back, either because it sees the task being done by others or because of an implicit assumption of reversibility. At first, even when the child has learnt which piece belongs in which recess there is no understanding of the need to line up the boundaries, so there is futile pressing.

Later the child may succeed by chance, using nearly random movements, but the probability of success with random movements is **very** low. (Why?)



Memorising the position and orientation **with great accuracy** will allow toddlers to succeed: but there is no evidence that they have sufficiently precise memories or motor control. Eventually a child understands that unless the boundaries are lined up the puzzle piece **cannot** be inserted. Likewise she learns how to place shaped cups so that one goes inside another or one stacks rigidly on another.

These changes require the child to build a richer ontology for representing objects, states and processes in the environment, and that ontology is used in a mental simulation capability. **HOW?**

Stacking cups are easier partly because of symmetry, partly because of sloping sides: both reduce the uniqueness of required actions, so the cups need less precision and are easier to manage.

Learning ontologies is a discontinuous process

- The process of extending competence is not continuous (like growing taller or stronger).
- The child has to learn about **new kinds** of
 - objects,
 - properties,
 - relations,
 - process structures,
 - constraints,...
- and these are different for
 - rigid objects,
 - flexible objects,
 - stretchable objects,
 - liquids,
 - sand,
 - mud,
 - treacle,
 - plasticine,
 - pieces of string,
 - sheets of paper,
 - construction kit components in Lego, Meccano, Tinkertoy, electronic kits...

I don't know how many different things of this sort have to be learnt, but it is easy to come up with many significantly different examples.

New Theory Vision Slide 39 Last revised: February 17, 2007 Page 53

CONJECTURE

In the first five years

- a child learns to run at least hundreds,
- possibly thousands, of different sorts of simulations,
- using different ontologies
 - with different materials, objects, properties, relationships, constraints, causal interactions.
- and throughout this learning, **perceptual capabilities are extended by adding new sub-systems to the visual architecture, including new simulation capabilities**

Some more examples are available in

<http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0601>

COSY-DP-0601 Orthogonal Competences Acquired by Altricial Species (Blanket, string and plywood).

New Theory Vision Slide 40 Last revised: February 17, 2007 Page 54

14 Much of what is learnt is about kinds of stuff

Human children (and presumably also chimpanzees, nest building-birds and members of other altricial species) learn many things about the environment by playful exploration, using a collection of special-purpose mechanisms developed by evolution for the task.

Part of what they learn concerns **the behaviour of various kinds of physical stuff** in the environment, including

- kinds of material like:
 - sand, water, mud, straw, leaves, wood, rock,
 - and in our culture also: things like paper, cloth, cotton-wool, plastic, aluminium foil, butter, treacle, velcro, meal, concrete, glue, mortar,
 - various kinds of food (meat, fish, vegetable matter, peanut-butter, etc.)
- kinds of components that can be combined to form larger objects including:
 - lego, meccano, tinker-toy, Fischer-technik, and many more,
 - including, for nest-building birds, twigs, leaves, etc.

‘Behaviour’ of such things includes their responses to being folded, crushed, picked up, thrown, twisted, chewed, sucked, pressed together, compressed, stretched, dropped, and also the properties of larger wholes containing them.

The variety of kinds of stuff and kinds of behaviour should not be thought of as a **continuum**, e.g. something that might be form a vector space parametrised by a collection of real-valued parameters. Rather there are qualitative and structural differences important in many sub-ontologies that have to be learnt separately (even if some precocial species have precompiled subsets).

A few examples follow: you can probably think of many more.

Cloth and Paper



You have probably learnt many subtle things unconsciously about the different sorts of materials you interact with (e.g. sheets of cloth, paper, cardboard, clingfilm, rubber, plywood).

That includes learning ways in which you can and cannot distort their shape.

Lifting a handkerchief by its corner produces very different results from lifting a sheet of printer paper by its corner – and even if I had ironed the handkerchief first (what a waste of time) it would not have behaved like paper.

Most people cannot simulate the **precise** behaviours of such materials but we can impose constraints on our simulations that enable us to deduce consequences.

In some cases the differences between paper and cloth will not affect the answer to a question, e.g. the example on the slide about folding a sheet of paper, below.

What do you know about cloth and paper?

There are probably many things you know about cloth and (printer) paper that you have never thought about, but implicitly assume in your reasoning about them, including imagining consequences of various sorts of actions.

Common features

- Both have two 2-D surfaces, one on each side.
- Both have bounding edges.
- Both can be made to lie (approximately) flat on a flat surface.
- Both can be smoothly pressed against a cylindrical or conical surface, but not a spherical (concave or convex surface)
- To a first approximation neither is stretchable, in the sense that between any points P1 and P2 there is a maximum distance that can be produced between P1 and P2, if there is no cutting or tearing.
- Both can be cut, torn, folded, crumpled into a ball....

Differences

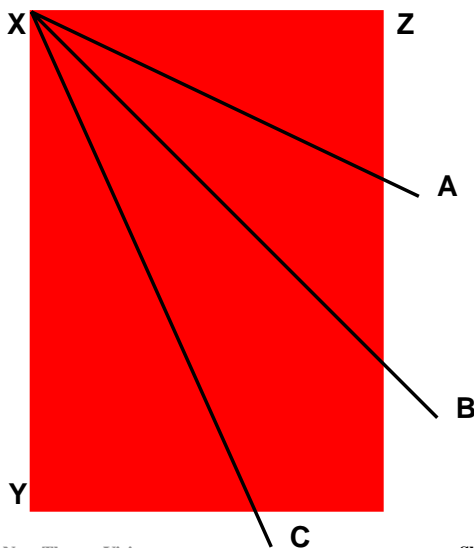
- most cloth can be slightly stretched (though some is very stretchy)
- Paper folded and creased tends to retain its fold, cloth often doesn't (there are exceptions, especially if heat is applied).
- Paper folded and not creased tends to return to its flatter state. It is more elastic.
- Paper folded once can stand upright resting on either a V-shaped edge or a pair of parallel edges.
- Paper is rigid within its plane (three collinear points remain collinear while the paper lies flat).

NOTE: tissue paper is somewhere in between.

Simulating folding of a sheet of paper

You can easily imagine folding a piece of red paper. Assume that it is ordinary paper, not a stretchy sheet of rubber.

What will happen to the corner Z if the sheet is folded along one of the lines A, B, and C, while the edge XY and adjacent portion of the sheet remains flat on the table.



Will the corner Z end up inside the red rectangle, outside it or on the edge of it?

You can probably think about this in several different ways (especially if you are a mathematician).

Some ways of thinking about it involve simulating the process of folding.

Others involve visualising or reasoning about where the moved edges will end up.

A very young child cannot do this, but eventually most will learn to think about folding of paper, and to see the effects of folds as **determined by** the structure, the nature of the material (paper) and the folding process.

Other simulations use different constraints.

15 Viewpoint matters - some viewpoints are 'vicarious'

The importance of viewpoint is obvious for any animal that moves, for self-motion can change the appearance of objects in a manner than depends on the shape of the object, its material, the lighting, the type of motion and what else is in the environment (actual or potential occluders).

What is not so obvious is that a part of the body, e.g. a grasping hand, may have a 'viewpoint' that is different from the visual viewpoint and which changes differently, as the hand moves or as something in the environment moves. E.g. something moving can block the eye's view of an object while leaving the hand's 'view' (route to the object) intact, and vice versa.

Likewise another person (or a child that needs help) may have a different and changing viewpoint.

So an intelligent animal or robot may need to be able to construct and reason about, or simulate properties of, 'vicarious viewpoints', i.e. viewpoints for others.

Contributors to simulation features

- We have so far seen that both shape and material can contribute to features of a simulation, including the constraints on what can and cannot change and what the consequences of change are.
- Another thing that can be important is **viewpoint**.
E.g. viewpoint can interact with opacity of materials, as well as with the mathematics of projection from 3-D to 2-D.

Sometimes a simulation includes a viewpoint

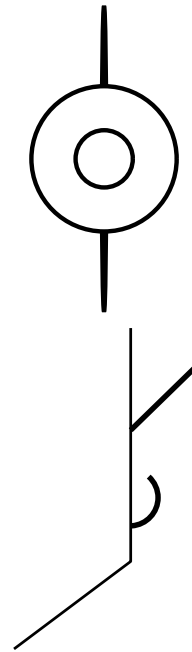
Doodles illustrate our ability to generate a simulation (possibly of a static scene) from limited sensory information (sometimes requiring an additional cue, such as a phrase ('Mexican riding a bicycle', or 'Soldier with rifle taking his dog for a walk').

In both of these two cases the perceiver is implicitly involved: one involves a perceiver looking down from above the cycling person, whereas the other involves the perceiver looking approximately horizontally at a corner of a wall or building.

In both cases the interpretation includes not only what is seen but also occluded objects: the simulation depends on knowing about opacity.

This does not imply that we have opaque objects in our brains: merely that opacity is one of the things that can play a role in the simulations, just as rigidity and impenetrability can.

The general idea may or may not be innate, but creative exploration is required to learn about the details.



We can see things from more than one viewpoint

- **Vicarious affordances:** a parent watching a child needs to be able to see what is and is not possible in relation to the child's needs, actions, possible intentions, etc. (It is also useful to be able to perceive a potential predator's affordances.)
- This may include such things as visualising the scene from the child's viewpoint, including working out what the child can and cannot see – and the possible consequences of the child seeing some things and not seeing others.
- Some people can draw pictures of how things look from some other place than their current location.
- This ability to contemplate the world from multiple viewpoints, not just one's own current viewpoint, is essential for planning, since at some future state in the plan one's location and orientation could be very different from what it is now, yet it still needs to be reasoned about in extending the plan.
- The ability to perceive and use information about 'vicarious' affordances (affordances for others) and the ability to perceive affordances for oneself in the past (e.g. thinking about a missed opportunity) or future (planning to use opportunities that have yet to be created) may use the same mechanisms **because both are disconnected from current viewpoint.**

Could that be the main point of substance behind all the fuss about "mirror neurons"? They should have been called abstraction neurons.

Seeing things from the viewpoint of your hand

The importance of hand-eye uncoordination!

- The evolution of body-parts for manipulation that can move independently of a major sensor perceiving what's happening (hands vs beak or mouth) had profound implications for processing requirements.
- Most animals are restricted to doing most of their manipulation with a mouth or beak, which cannot move much without the eyes moving too.
- If your eyes move as your gripper moves, because they are closely physically connected, then the sensory-motor contingencies linking actions and their sensory consequences will have strong, useful regularities that can be learnt and used.
- If a gripper can move independently of the eyes then the variety of relationships between actions and sensed consequences explodes.
The explosion can be reduced by modeling action at a level of abstraction removed from sensory changes: e.g. by representing actions as altering 3-D structures and processes (including subsequent actions), independently of how they are sensed.
- The mapping between sensory data and what is perceived becomes very indirect, and there may need to be several intermediate layers of interpretation: perception becomes akin to constructing a structured theory to explain complex data. (Compare the 'dotty picture' example, above.)
- **This is one of many reasons for NOT regarding perception as simply concerned with detecting sensory-motor contingencies.**

Sensory-motor vs action-consequence contingencies

Two evolutionary 'gestalt switches'?

The preceding discussion implies that during biological evolution there was a switch (perhaps more than once) from

insect-like understanding of the environment in terms of **sensory-motor contingencies** linking internal motor signals and internal sensor states (subject to prior conditions),

to

a more 'objective' understanding of the environment in terms of **action-consequence contingencies** linking changes in the environment to consequences in the environment,

followed by

a further development that allowed a **generative** representation of the principles underlying those contingencies, so that novel examples could be predicted and understood, instead of everything having to be based on statistical extrapolation.

To be more precise, it was an **addition** of a new competence rather than a **switch**

One of the major drivers for this development could be evolution of body parts other than the mouth that could manipulate objects and be seen to do so.

However the cognitive developments were not **inevitable** consequences: e.g. crabs that use their claws to put food in their mouth do not necessarily use the more abstract representation.

16 No good theories about shape perception exist

A huge amount of work on machine vision totally ignores shape and is concerned only with recognition, classification, prediction, or tracking, more or less treating the world as two-dimensional.

However there are some attempts to get machines to perceive shape.

Unfortunately these mostly seem to use inadequate requirements for shape perception. E.g. using vision and laser-scanning or whatever, to produce a detailed 3-D model of space occupancy which can be given to computer graphics programs to project images from any viewpoint in different lighting conditions may be very useful for many applications (e.g. medical imaging, and computer games) this does not give the computer a kind of understanding of shape that is required for manipulating objects.

Perception of shape is not shape-reconstruction

What sort of 3-D interpretation is required depends on what it is to be used for.

Shape perception in computers is often demonstrated by giving the machine one or more images, from which it constructs a point-by point 3-D model of the visible surfaces of objects in the scene.

This achievement is then demonstrated by projecting images of the scene from new viewpoints.

But there is no evidence that any animal can do that and very few humans (e.g. some artists) can produce accurate pictures of viewed objects using a new viewpoint, whereas many graphics engines do it.

Human/animal understanding of shape, including having information relevant to action and prediction, is very different from having a point by point 3-D model

The point of perception is not making images: the results must be useful for action – e.g. building nests from twigs, peeling and dismembering food in order to get at edible parts, escaping from a predator, making a tool, using a tool.

A 'percept' constructed by the perceiver needs to include information about what is happening, what could happen and what obstructions there are to various kinds of happening (positive and negative affordances).

These happenings are of many different kinds, so different kinds of information must be synthesised from sensory information (influenced by prior knowledge, prior ontologies, prior goals).

How is it possible?

- It has long been known that the problem is too unconstrained to be solvable – every 2-D image is inherently capable of being generated by infinitely many 3-D scenes.
- It has long been conjectured that the environment is constrained in ways that make the problem **contingently** solvable — where some constraints may be learnt by the individual perceiver and others are derived from the genetically determined structure, functions, and processing mechanisms of perceptual systems: **The ‘cognitively friendly environment’ hypothesis.**
- Examples include use of binocular vision (which helps only a little, and only at short distances), motion perception (which can be far more important, whether the motion is in the perceiver or in the perceived objects), and assumptions about the nature of various materials, e.g. how rigid they are, what their surface texture is, the kinds of lighting found in various situations, knowledge of the effects of occluding opaque objects, intervening shrubbery, distortions caused by heat-haze, etc.
- Of course, we and other animals are not perfect perceivers and banking on these constraints can sometimes lead us into error (e.g. the Ames room and other illusions, including some used by animals, such as camouflage) though usually the implicit assumption of cognitive friendliness works well.

How can brains do all this?

What good are examples without any theory of how brains do it?

- **Beware: if we theorise on the basis of too few kinds of examples we may come up with inadequate theories: a common problem in AI, philosophy, psychology and neuroscience.**
- If all the above is correct, human brains need to be able to run very many different kinds of simulations:
 - including processes involving stones, blocks, string, paper, sand, cloth, mud, plasticine, rigid materials, flexible materials, materials that are rigid in two dimensions and flexible in one (e.g. paper), water, sand, mud, cotton wool, plasticine, wire, fibrous materials, viscous liquids, various kinds of meat, various kinds of vegetable matter, brittle materials, stretchable materials, thin films, solid lumps of matter, and many more.
- We are not restricted to simulating what has occurred in our evolutionary history: children can learn to play with and think about toys and devices none of their ancestors ever encountered – e.g. skipping ropes, slinky springs, zip fasteners, velcro, scotch tape, computer games and future inventions too.

If we start building explanatory models based on too few explananda we may fool ourselves into accepting inadequate theories.

So we should seek a ‘generative’ explanation. That’s an old idea, but if the generative explanation is too simple (like current popular theories of learning) it may work on toy examples but fail hopelessly in the tasks summarised here.

17 Not only humans

If we try to find out more about what different sorts of animals can and cannot do, that may help us to understand the evolution of human competences of the sorts described here, and thereby give us clues as to the mechanisms involved,

Finding fracture lines between different subsets of competences can help us notice important features of those competences that have implications for different ontologies, different forms of representation, different mechanisms, different architectures, different kinds of learning, etc.

It may even be useful to regard young human children as if they were members of different species and not just assume that they are smaller versions of human adults.

We should try to find out in great detail what different infants and toddlers can and cannot do and how many different routes there are through their epigenetic landscape (Waddington), and how the landscape (the set of possible developmental pathways) depends on the physical and cultural environment.

This may help to provide much stronger constraints on explanatory theories (of perception, learning, development, control of actions, etc.) than we have at present.

How many non-human species?

Betty the hook-making New Caledonian crow.

Give to google: betty crow hook:
You'll find a link to the oxford zoology lab, with videos of Betty making hooks in different ways.

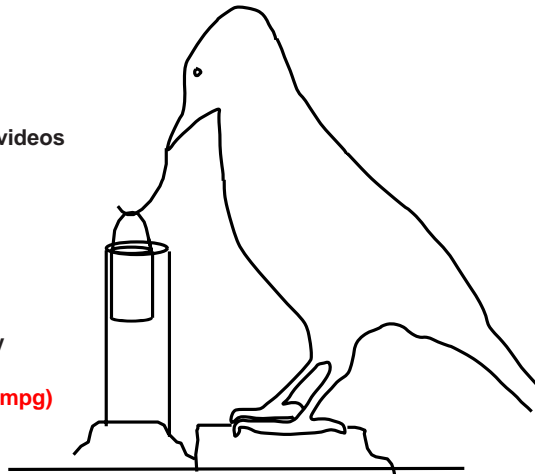
She **appears** to be a Kantian causal reasoner.

See the video here:

<http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm>

Contrast the 18 month old child attempting unsuccessfully to join two parts of a toy train by bringing two rings together

(http://www.cs.bham.ac.uk/~axs/fig/josh34_0096.mpg)



Does Betty see the possibility of making a hook before she makes it?

She seems to. How?

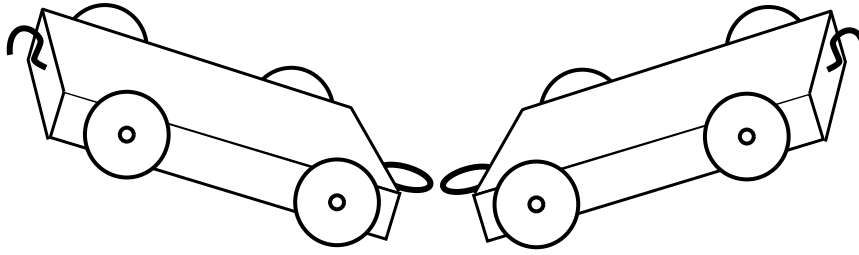
Understanding how hooks work

- Betty seems to understand how hooks work when she uses hooks to lift a basket of food out of the glass tube.
- The depth of understanding seems even greater when she demonstrates her ability to make hooks from straight pieces of wire in several different ways. I have also seen her make a hook from a long thin flat strip of metal.
- The behaviour is clearly not random trial and error learning behaviour: she seems to know exactly what to do, even though she does things in slightly different ways, e.g. making hooks using different techniques.
- Note that in Betty's environment far more distinct motions are possible than in the multi-rod linkage a few slides back: how does she confidently select a course through the continuum of continua?
The answer cannot simply be: by running a simulation, because the simulation might have the same problem of under-determination.
- A young child does not start off understanding how a hook and a ring can interact in such a way as to allow the hook to pull the ring and what it is attached to.
- At some stage that (Kantian) understanding develops.
But I don't think anyone knows how – even if some psychologists know when.
- The next slide points to a video showing a child who has not yet got there.

18 A child can appear less competent than a crow

We next show a video of a 19 month old child who is competent in many ways but seems to fail to understand how a hook and ring are used to join up a toy train.

Defeating a 19 Month old child



See the movie of an 19-month old child failing to work out how to join up the toy train – despite a lot of visual and manipulative competence also shown in the movie.

- http://www.jonathans.me.uk/josh/movies/josh34_0096.mpg
4.2Mbytes

- http://www.jonathans.me.uk/josh/movies/josh34_0096_big.mpg
11 Mbytes

The date is June 2003, when he was 19 months old. (Born 22 Nov 2001)

A few weeks later he had no problem joining up the train.

Was he a Humean causal learner or a Kantian causal learner?

I suspect the latter, but specifying the simulation model developed by a learner who understands hooks and rings will not be easy.

New Theory Vision Slide 54 Last revised: February 17, 2007 Page 73

19 Running 2-D or 3-D simulations to answer questions

Perhaps the child who fails to join up the train does not understand because he has not yet learnt to simulate processes in which a hook and a ring form a connection that is useful for pulling.

Why not? Why are some competences innate, and some learnt. Why are some learnt very early and some only later.

Maybe we still have to understand the dependency relations between hundreds, or thousands, of sub-competences.

There are many problems we can solve, by running 2-D or 3-D simulations.

Some examples follow.

We can run a simulation to answer a question

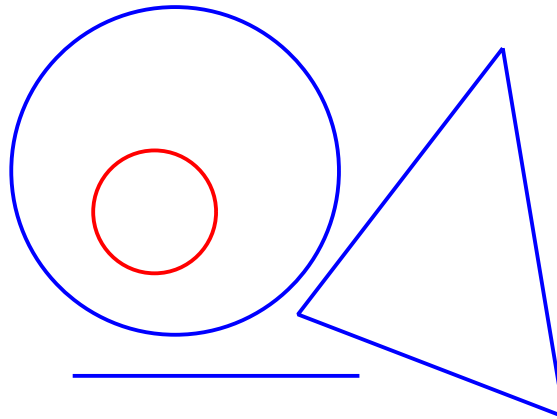
As the horizontal blue line moves vertically upwards, it may temporarily intersect one or more other lines, in one or more places.

Which intersection points will appear as it moves and how will their locations change as the motion continues?

What is the largest number of intersections that will co-exist as the line moves up?

Our simulations can create new entities and relations, and we can count entities created.

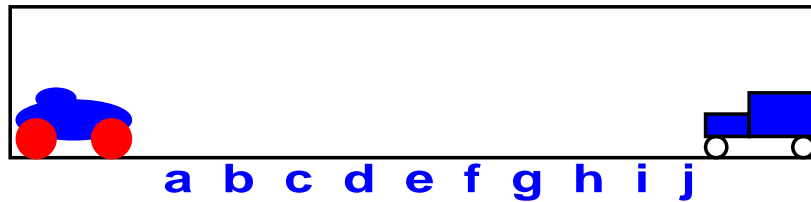
Some people can do this in their heads, whilst others may need to slide a ruler up the page: but they are using different mechanisms (physical machines or virtual machines) to do the same thing. As my 1971 paper pointed out.



NOTE:

A mathematician can answer by reasoning about the simulation, instead of running it, after noticing what kinds of discrete transitions can occur, e.g. the end of a line entering or leaving a closed region.

Simulating potentially colliding cars



The two vehicles start moving towards one another at the same time.

The racing car on the left moves much faster than the truck on the right.

Whereabouts will they meet – more to the left or to the right, or in the middle?

Where do you think a five year old will say they meet?

Five year old spatial reasoning



The two vehicles start moving towards one another at the same time.

The racing car on the left moves much faster than the truck on the right.

Whereabouts will they meet – more to the left or to the right, or in the middle?

Where do you think a five year old will say they meet?

One five year old answered by pointing to a location near 'b'

Me: Why?

Child: It's going faster so it will get there sooner.

What is missing?

- Knowledge?
- Appropriate representations?
- Procedures?
- Appropriate control mechanisms in the architecture?
- A buggy mechanism for simulating objects moving at different speeds?

Mr Bean's underpants

This paper (from a conference on thinking with diagrams in 1998)

<http://www.cs.bham.ac.uk/research/cogaff/00-02.html#58>

discusses how we can reason about whether Mr Bean (the movie star) can remove his underpants without removing his trousers.

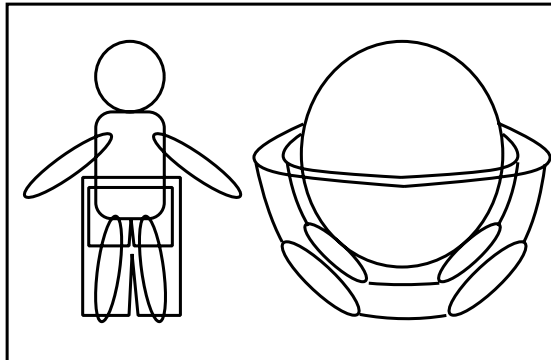
People often don't see all the possibilities at first.

The paper discusses how changing the simulation to a topologically 'equivalent' one can help us count the possible ways to perform the task.

Children can learn to perform such actions (as party tricks) physically long before they can reason with the mental simulations.

What changes as the simulation ability develops?

In part it seems to require an introspective ability to understand the nature of the simulations we use.



See

Jean Sauvy & Simonne Sauvy *The Child's Discovery of Space, From Hopscotch to Mazes: an Introduction to Intuitive Topology* (Translated P.Wells 1974).

New Theory Vision

Slide 58

Last revised: February 17, 2007 Page 79

21 Seeing structure, motion, and invariants in mathematics

Hume thought that all knowledge was either **analytic** (i.e. true by definition and essentially empty), or **empirical**, requiring experiment and observation for its confirmation, and therefore capable of turning out false in new situations.

Kant thought there were counterexamples, especially in mathematical knowledge, which he claimed was **synthetic**, i.e. amplified our knowledge, and **non-empirical** (or *a priori*), i.e. immune from empirical refutation.

My Oxford D.Phil thesis (completed in 1962, but never published) was an attempt to defend Kant against Hume, but, like Kant, I did not have adequate conceptual tools for the job. We are a little closer now insofar as we may soon be able to design working models of how a mathematician uses mechanisms that are needed for perception of and thinking about complex structures can be deployed in making mathematical discoveries, including seeing why $7 + 5$ must always be 12 (Kant's example).

The suggestion that follows is that this is connected with our understanding invariant properties of one to one mappings, which most people can visualise in terms of spatial connection, even though the mathematical notion is far more general and not restricted to spatial objects.

A child learning to count eventually has to understand all this, in order to understand what numbers (at least the positive integers treated as cardinal numbers) are and what mathematical truths are. Unfortunately their teachers may be too confused to help children who do not discover these things spontaneously.

When we go beyond the positive integers things get far more complex in ways that very few people understand, alas, so they just learn rules of thumb that work – their minds remain partly underdeveloped for life. (This is true of all of us, in some respects.)

KANT'S EXAMPLE: $7 + 5 = 12$

Kant claimed that learning that $7 + 5 = 12$ involved acquiring *synthetic* (i.e. not just definitionally true) information that was also not *empirical*. I think his idea was related to the simulation theory of perception – but I am guessing.

You may find it obvious that the equivalence below is preserved if you spatially rearrange the twelve blobs within their groups:

$$\begin{array}{r} \text{ooo} \\ \text{ooo} \\ \text{o} \end{array} + \begin{array}{r} \text{o} \\ \text{o} \\ \text{ooo} \end{array} = \begin{array}{r} \text{oooo} \\ \text{oooo} \\ \text{oooo} \end{array}$$

Or is it?

How can it be obvious?

Can you see such a general fact?

How?

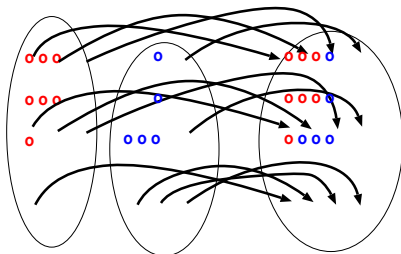
What sort of equivalence are we talking about?

I.e. what does “=” mean here?

Obviously we have to grasp the notion of a “one to one mapping”.

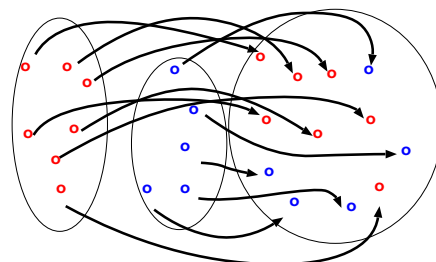
That **can** be defined logically, but the idea can also be understood by people who do not yet grasp the logical apparatus required to define the notion of a bijection — if they have a way of thinking about the consequences of motion of the blobs.

SEEING that $7 + 5 = 12$



Then rearrange the items, leaving the strings attached.

Is it 'obvious' that the correspondence defined by the strings will be preserved even if the strings get tangled by the rearrangement?



Join up corresponding items with imaginary strings.

Is it 'obvious' that the same mode of reasoning will also work for other additions, e.g. $777 + 555 = 1332$

Humans seem to have a 'meta-level' capability that enables us to understand why the answer is 'yes'. This depends on having a model of how our model works – e.g. what changes and does not change if you add another pair of objects joined by a string.

But that's a topic for another occasion.

22 High level perceptual processes can ignore low-level details

I am suggesting that when we watch or imagine things moving we simulate the motion (i.e. we create and run representations) at different levels of abstraction.

Some of them we probably never become conscious of as they are used only in relatively automatic control of common processes, for instance as optical flow patterns are used in posture control.

What we say we are **conscious of** is often closely related to what we can **report**, to ourselves or to others, and that will typically be things happening at a high level of abstraction, that are relevant to our current goals and needs, though we can direct our attention to details just for the sake of examining details, and we can become aware of details that are too rich and complex to be reported, even to ourselves, e.g. watching swirling rapids in a fast flowing river or hundreds of leaves stirring in the wind.

What we are conscious of seeing may depend on what the current task is, and sometimes we do not notice details even if a low level system processes them – e.g. because what we attend to when answering a question includes only the contents of the more abstract simulations.

But that does not mean that the details have not been processed, as the next example shows.

A well known example of controlled hallucination



In this case some people only see an abstraction – a familiar phrase, rather than what is actually visible in the circle.

Similarly when we run simulations we may sometimes hallucinate what we expect to be in the environment rather than what is actually there.

Do you see only a familiar phrase? If so, read on.

A part of you may see what 'you' do not see!

Often people who have been shown the example on the previous slide and are convinced, even after insistent questioning, that what they see is just a familiar phrase, can be made to realise their mistake, even with their eyes shut.

- Ask subjects who claim to have seen only 'PARIS IN THE SPRING' to shut their eyes.
- Then ask **one** of these two questions
 - How many words were in the circle?
 - Where was the 'THE'?
- Some of them realise, even with their eyes shut, that what they were certain they had seen was not what they had actually seen.

This seems to show that, at least for such a person, it is wrong to ask 'What did he/she see?', for the answer will be different for different **parts** of the person.

A part of you may record the layout of the words in the circle even though another part (central to social interactions) decides that it is a familiar phrase on the basis of evidence that is often perfectly adequate, and it does not check for consistency with the low level detail.

In a cognitively 'friendly environment' (assumed for Popeye) where decisions sometimes have to be taken quickly, this could be a good design, even if it occasionally causes errors.

Learning when to be more thorough can be useful in some environments!

This idea may explain phenomena revealed in experiments on 'change blindness' – where experimenters wrongly assume that we know what we see, whereas much perception is subconscious.

New Theory Vision Slide 62 Last revised: February 17, 2007 Page 85

Seeing non-existent motion

There are many optical illusions in which things appear to be moving when they are not, including motion after-effects, and patterns used in so-called 'op-art'.

See <http://www.michaelbach.de/ot/index.html>

Nothing I have said explains any particular phenomenon of illusory motion, but the existence of such things is perhaps less surprising if we think of all visual perception as involving the running of process simulations controlled in part by sensory data, and subject to **presumed** constraints that may sometimes be inferred wrongly.

If all that powerful apparatus exists ready to be used at very short notice, it may easily be triggered into action by a variety of partial cues: some erroneous interpretations are very likely in that case — but in a 'cognitively friendly' environment the result will be fast decisions that are mostly correct.

In relatively simple cases we can take in all the relevant structure and work out what must be happening: this is the basis of mathematical capability.

New Theory Vision Slide 63 Last revised: February 17, 2007 Page 86

23 What the simulation theory does and does not say

So far I have given many examples, and talked very vaguely about perception and reasoning as involving various kinds of simulations, using different ontologies with different sorts of constraints, different viewpoints, etc.

But the theory is easily misunderstood – and also still has many gaps.

I'll now try to make it a little more precise, including saying what I am NOT claiming.

The concurrent simulation theory in more detail

- Different simulations of the same scene may be used in different sub-mechanisms running simulations at different levels of abstraction and serving different functions.
- Some parts of simulations may **go beyond sensory data**, e.g. including unobserved sub-mechanisms (Kant)
- Some of the processes are **continuous** some **discrete**.
- The continuous and discrete processes may both have **different levels of resolution**.
- There may be **gaps** in the simulation at all levels (for different reasons)
- **Mode of processing can change dynamically**: parts of the simulation may be selected for more detailed processing, or type of processing can be changed.
- Seeing static scenes involves running **simulations in which nothing happens** – though many things could happen (cf. seeing affordances).
- The mechanisms originally evolved to support perceptual and motor control processes but became detachable from that role in humans and can be used to think about things that could never be observed,
e.g. search spaces, high-dimensional spaces, infinite sets, including operations on transfinite ordinals (move all the odd numbers after the even numbers and reverse their order).
See my paper 'Diagrams in the mind' 1998
<http://www.cs.bham.ac.uk/research/cogaff/96-99.html#38>

Development of perceptual sub-systems

The ability to run these simulations is not static, and may not even exist at birth:

- Visual capabilities described here develop in part on the basis of developing architectures for concurrent simulations and in part on the basis of learning new types of simulation, with appropriate new ontologies and new forms of representation.
- The initial mechanisms that make all of this possible must be genetically determined (and there may be limitations caused by genetic defects).
- But the *contents* of the abilities acquired through various kinds of learning are heavily dependent on the environment – physical and social, and on the individual's history. Some innate content is needed for bootstrapping.
- For instance someone expert at chess or Go will see (slow-moving!) processes in those games that novices do not see.
- Expert judges of gymnastic or ice-skating performance will see details that others do not see.
- An expert bird-watcher will recognize a type of bird flying in the distance from the pattern of its motion without being able to see colouring and shape details normally used for identification.

A deeper theory would explain the variety of types of changes involved in such developments: including changes in ontologies used, in forms of representation, and perhaps also in processing architectures.

These will be changes in virtual machines implemented in physical brains.

Seeing intentional actions

Seeing a person or animal or machine doing something may involve a richer ontology than is required for seeing physical things moving under the control of purposeless physical forces.

- If you see a marble rolling down a slope occasionally changing direction or bouncing into the air as a result of surface irregularities or stones in its path, your simulation may include changes of position, speed and direction of motion, all consistent with what you know about physical objects.
- If you see a person walking down a slope occasionally moving to one side and picking things off bushes, you will see not only physical motion, but **the execution of an intention**, possibly several intentions, e.g. getting to something at the bottom of the slope, collecting biological specimens, and eating berries.
- One of the things a child has to learn to do is interpret perceived motion in terms of inferred goals, plans and processes of plan execution. Thus the simulations run when intentional actions are perceived may include a level of abstraction involving **plan execution**.

For a recent discussion see Sharon Wood, 'Representation and purposeful autonomous agents' *Robotics and Autonomous Systems* 51 (2005) 217-228

<http://www.cogs.susx.ac.uk/users/sharonw/papers/RAS04.pdf>

- When several individuals are involved, there may be several concurrent, interacting, processes with different intentions and plans to simulate. Learning to understand stories beyond the simplest sequential narratives requires learning to do this. (Contrast coping with 'flashbacks'.)

24 Seeing structures, processes and causation

A separate, partly overlapping, paper deals with how all this is relevant to understanding causality. Here we merely make the claim of relevance.

Conjecture

A great deal of our understanding of causality is intimately bound up with our ability to create constrained, deterministic simulations, and to learn about their properties by ‘playing’ with them.

We are not born with all the specific simulation capabilities we have, but we, and possibly several other altricial species, are born with mechanisms for developing such simulations — depending on what is encountered in the environment.

We are born equipped to become Kantian causal reasoners about more and more aspects of the environment, though there are always residual unexplained but useful correlations.

Similar remarks can be made about the history of science and technology.

See <http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0506>

25 What I am NOT saying

The theory being proposed is easily misinterpreted.

The following slides attempt to explain what is **not** being said, by pointing out that some tempting interpretations of the theory are wrong.

Disclaimers: No claim is made:

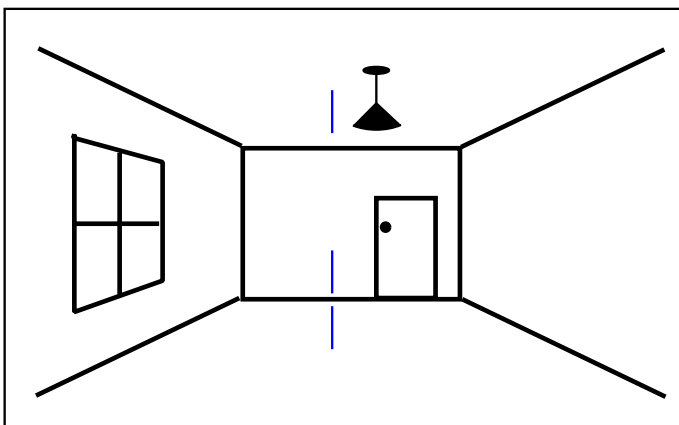
- That the simulations at any level are complete
- That they are accurate (errors, imprecision and fuzziness abound)
- That we are aware of all the simulations we are running
- That only humans can do this
- That all humans can run the same kinds of simulations
 - Different kinds of education, different kinds of training, e.g. artistic, athletic, mathematical training, playing with different kinds of toys, etc. can all produce different ontologies, representations and simulation capabilities. Even children with similar competences may get there via different routes along a partially ordered network of trajectories. **There are genetic differences too – e.g. ‘Williams syndrome’ children don’t develop normal spatial competences.**
- That it is obvious how to implement these ideas in artificial visual systems
- That the theory is compatible with any current theory of learning
- That the theory is compatible with known brain mechanisms
 - We may have to search for previously undiscovered mechanisms (including previously unknown types of virtual machines implemented in brains)
 - See Trehub’s book (*The Cognitive Brain, 1991*) for some relevant ideas.
 - There are probably lots of things I should have read but have not.
 - There is considerable overlap with the BBS paper by R.Grush (2004): The Emulation Theory of Representation.

Isomorphism is not needed

Here's a modified version of a picture from chapter 7 of *The Computer Revolution in Philosophy*, also in the 1971 IJCAI paper.

Objects and relations within a picture need not correspond 1 to 1 with objects and relations within the scene, as is obvious from 2-D pictures of 3-D scenes.

For example: pairs of points in the image that are the same distance apart in the image can represent pairs of points that are different distances apart in 3-D space – e.g. vertically separated points on the walls, and horizontally separated points on the floor and ceiling. (And *vice versa*.)



Some pairs of parallel edges in the scene are represented by parallel picture lines, others by converging picture lines.

The small blue lines can be interpreted in different ways, with different spatial locations, orientations and relationships. On each interpretation the structure of the image remains unchanged, but the structure of the 3-D scene changes.

MAJOR DISCLAIMER

I am not claiming that simulations have to be isomorphic with what they simulate

- As pointed out in my 1971 paper, analogical representations use relations to represent relations but they need not be **the same** relations:
Think of a 2-D picture of a 3-D scene (the same 2-D relation 'above' can correspond to different 3-D relations in different parts of the picture – floor, far wall, ceiling).
See <http://www.cs.bham.ac.uk/research/cogaff/crp/chap7.html>
- Not all simulations of spatial processes have to be spatial: it may often be simpler to use equations, for example, and psychological behavioural experiments may be wholly unable to determine which kind of implementation is used without having access to design information.
- Somehow we have developed enormously flexible ways of using mappings between one changing structure and another changing **or static** structure – it is a matter of learning what kinds of formalism with what kinds of constraints do and do not work for particular tasks.
E.g. programming language constructs can map onto dynamic graphical displays.
- The ability I am talking about goes on being developed throughout life as we acquire more and more kinds of expertise.
- **That means a complete theory will have to explain that acquisition process – and no finite theory will explain all past, present and future human competence.**

Disclaimer: thinking is not simulated hearing

I am not claiming that thinking depends on linguistic competence combined with simulation.

- There are people who believe that since much of thinking seems like talking to yourself, mental processes involve simulated talking.
- However there is an argument that most of the features of human language (including recursive syntactic structure and compositional semantics) had to exist prior to the inter-personal use of language, in order to support many pre-linguistic competences (such as perception, planning and way-finding) and also as a basis for the manipulation of structures involved in generating or understanding language.

The argument was presented in 1979 in 'The primacy of non-communicative language' available here

<http://www.cs.bham.ac.uk/research/cogaff/81-95.html#43>

- From this viewpoint the ability to run abstract simulations does not depend on linguistic competence, but rather linguistic competence arises out of this earlier representational competence that is present in children while they are learning to talk, and may exist in other animals that never learn to use an external language like ours.

Conjecture: evolution of what we call use of language depended on externalisation of pre-existing simulations using mechanisms that evolved prior to the development of language.

26 Some inadequacies of closely related theories

The theory being developed here was not pulled out of a hat.

There were many prior influences including theories that emphasised aspects of the current theory.

We can help to clarify the theory by identifying gaps and limitations in other theories, which, it is claimed, are addressed by the theory presented here, even though the work is not yet complete.

A full analysis of strengths and weaknesses of previous theories would require several books. Here I merely offer a few pointers.

Inadequate alternative theories

Among the precursors to the theory are several that in different ways are inadequate, despite providing useful steps in the right direction.

- One general kind of inadequate theory assumes that what is perceived can be expressed as a collection of measures, sometimes called 'state variables', (e.g. coordinates, orientations, and velocities of objects in the scene) and that what is simulated can be expressed as continuous or discrete changes in a (possibly) large vector of state variables.
- This kind of numerical representation is inadequate because it fails to capture **the structure** of the environment, e.g. the decomposition into objects with parts, and with different sorts of relationships between objects, between parts within an object, between parts of different objects, etc.
People who are familiar with a particular collection of mathematical techniques keep trying to apply them everywhere instead of analysing the problems to find out what forms of representation are really required for the tasks in hand.
- Many theories do not do justice to the diversity of functions of vision. E.g. some people seem to think the sole or main function of vision is recognition of instances of object types.
- Most theories of vision do not allow that we see not only what exists but what can and cannot happen in a given situation – affordances.
- Dynamical systems theorists have some of the right ideas but restrict ontologies and forms of representation to what physicists understand.

Terminology

- Some people distinguish simulation, emulation, imagery, etc.
- What I call a simulation is a **representation of a process** that can be used for a variety of purposes, e.g. recording, predicting, tracking, explaining, controlling.
- A simulation may itself be a process, or it may in some cases be a re-usable static trace of a process, e.g. an executable plan, even a plan with loops and conditionals – with a ‘now’ pointer.
- The same process may be simulated at different levels of abstraction:
 - simulations run at a high level may be very much faster than what they represent.
- Different sorts of simulations are useful for different purposes.
- A child continually learns new sorts of simulations and new uses for old sorts.
- Some running simulations can change direction, can explore options.
- Some simulations are continuous, and some discrete, and some simulated processes are continuous and some discrete.
 - A continuous simulation may represent a discrete process and *vice versa*.
 - It is difficult for a continuous simulation do searching, e.g. in a space of possible explanations or possible plans: discretisation makes multi-step planning feasible.
- A simulation may change in complexity and structure as it runs (e.g. simulation of development of an embryo — unlike simulations that involve a fixed dimensional state vector).
- The things that change in a simulation need not be numerical variables.
- We probably don’t yet know all the powerful ways of representing processes that evolution may have discovered and implemented in brains.
- In principle a simulation can itself be simulated (e.g. at a higher level of abstraction) – as in John Barnden’s ATT-META system. <http://www.cs.bham.ac.uk/jab/ATT-Meta/>

28 Re-runnable check-points

One of the consequences of discretisation is support for multi-step deliberation, e.g. systematic searching for a plan, including use of back-tracking.

Re-runnable check-points

- When searching for a solution to a problem we often have to explore a branching space of possibilities.
- Continuous simulations are not good tools for exploratory searching because there are always infinitely many possible branch points with infinitely many branches.
- This can be overcome by doing the searching with the aid of a discrete, more abstract, symbolic version of the simulation, and saving check-points, which can later be compared with one another.
- Ideally the check-points should be able to generate new lower-level runs of the simulation, when you back-track to a check-point.
- But for this, fully fledged deliberative mechanisms (for exploring answers to 'what if questions') could not really use simulations.
- So the development of discrete (symbolic) forms of representation was a major step for evolution. It had profound consequences including making mathematics and human language possible.

Some animals probably use discrete symbols in internal languages.

<http://www.cs.bham.ac.uk/research/cogaff/81-95#43>

29 How the theory arose in the context of the CoSy project

The role of requirements analysis in theory construction and in design of complex systems is very important.

Unfortunately it is not always done well because people assume requirements are obvious and immediately start designing and building systems.

Or they take existing systems and work hard to improve them, without stepping back and asking: why is this needed, and what else is needed?

How the theory arose

One of the tasks in the CoSy project was to analyse requirements for representations used by a robot (the PlayMate) with the ability to manipulate 3-D objects on a table top.

<http://www.cs.bham.ac.uk/research/projects/cosy/>

<http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html>

While trying to understand requirements for a robot with manipulative capabilities watching 3-D processes in which one complex structured object moves in relation to another, I realised there were all sorts of implications, e.g. for my old interest in reasoning using analogical representations, the nature of mathematics, the understanding of causality, the role of affordances in perception, learning,

Why?

Because structured 3-D objects involve **multi-strand relationships** (parts are related to other parts of the same and different objects in the scene).

So moving 3-D objects can produce changes in multi-strand relationships: i.e. **multi-strand processes?**

How could that be represented? Answer: using simulations – but different sorts of simulations for different purposes, even when the physical scene is the same.

But there is still a huge amount of detailed work to be done, about forms of representation, mechanisms, architectures, development, learning, varieties of implementation....

See this CoSy technical report: 'Requirements study for representations':

<http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0507>

Finding requirements through deep analysis of tasks

Most 3-D scenes have multiple objects with multiple parts that stand in multiple relationships forming a richer web than we can express in language, though we can summarise salient aspects, e.g.

THE CAR MOVES INTO THE GARAGE

At every moment there are

- many parts of the car
- in multiple relations to one another
- many parts of the garage and things in the garage
- many relationships between parts of the car and parts of the garage (and other things in it),
- where parts and relationships exist at different levels of abstraction (physical, geometrical, topological, causal, functional, aesthetic, ...)

In short **what we can see, and think about, and learn about, and talk about, are multi-strand relationships, and when multi-strand relationships change we get multi-strand processes.**

Structures vs combinations of features

It is important to understand the difference between

- **Categorising**
- **Perceiving and understanding structure.**

You can see (at least some aspects of) the structure of an unfamiliar object that you do not recognise and cannot categorise: e.g. you probably cannot recognise or categorise this, though you see it clearly enough.

```
Oooo
Oooooo-----+
OOooooOOO    +
|oooOOoooo----+
+-----+
```

What is seeing without recognising?

There's a huge amount of work on visual recognition and labelling e.g. statistical pattern recognition.

But does that tell us anything about perception of structure?

Much work on vision in AI does not get beyond categorisation.

There is some work that attempts to identify structure from visual images, but the form in which structure is represented is merely a volumetric model, which may be very suitable for generating graphical displays from different viewpoints, but does not include any **understanding of the structure by the computer** – it leaves the main representational problems unsolved.

There is something even more subtle and complex than perception of structure.

New Theory Vision Slide 77 Last revised: February 17, 2007 Page 107

Perceiving structures vs perceiving affordances

Structures

things that exist, and have relationships, with parts that exist and have relationships

Affordances (positive and negative)

processes that could or could not (sometimes conditionally could or could not) be made to exist by the agent, with particular consequences for the perceiver's goals, preferences, likes, dislikes, etc.:

modal, as opposed to categorical, types of perception.

- Betty looks at a piece of wire and (maybe??) sees the possibility of a hook, with a collection of intervening states and processes involving future possible actions by Betty.
- The child looks at two parts of a toy train remembers the possibility of joining them, but fails to see the precise affordances and is mystified and frustrated: presumably he sees parts and structural relationships because he can grasp and manipulate them in many ways. But he appears not to see some affordances.
- Seeing affordances seems to be related to being able to run simulations of unseen but possible processes in registration with the scene.

How specialised are the innate mechanisms underlying the abilities to learn categories, perceive structures, understand affordances, especially structure-based affordances.

Millions of years of evolution were not wasted!

New Theory Vision Slide 78 Last revised: February 17, 2007 Page 108

Some tasks for a crow-challenging robot?

UPDATING THE BLOCKS WORLD

Using a two-finger gripper, what actions can get

from this:



to this:



and back again?

Or with saucer upside down?

Unfortunately even perceiving and representing the initial or final state (e.g. as something to copy) seems to be far beyond the capabilities of current AI vision systems, let alone thinking about possible actions to transform one to the other.

Some tasks for a crow-challenging robot? (2)

Consider how, prior to the action, the agent has to

- identify parts of objects, or parts of parts, e.g. the edge of the handle, or the far edge of the handle or a certain portion of the edge of the saucer
- see and understand their shapes and relationships
- identify possible actions: grasping **this thing here** from **this direction**
Could such deliberative premeditation use the action schema (operator) with approximate, qualitative parameters instead of the more definite actual parameters that would be used if the action were performed?
- think about various effects of actions, including changing effects of continuous processes

NOTE: there are problems here partly analogous to problems of reference and identification in language, except that the mode of reference is not linguistic and what is referred to typically cannot be expressed in language because it is anchored in non-shared structures and processes.

(Internal 'attention' processes are partly like external pointing processes: virtual fingers.)

Vision: the hardest problem

VISUAL INPUT WE SHOULD BE ABLE TO COPE WITH (FOR A MACHINE THAT CAN MANIPULATE OBJECTS):



We need to be able to see things we cannot easily express in language.

- This bit is concave. This bit is shiny. This bit is a reflection of the cup. **How is 'this bit' identified?**
(Compare 'virtual finger theories' – Z.Pylyshyn's FINST ?)
- As I move left and right those reflections move on that edge of the cup. **How are 'those' and 'that' identified?**
- This is how I should move my hand to grasp the spoon.
A different kind of 'this' (manner, route, method) – how identified?
Perhaps partially instantiated parameters in some operator?
- We need to understand ontologies and representations for active (possibly pre-linguistic) agents
- What details are 'visible' in image depends on how it is processed (Top picture, processed different ways gives middle or bottom picture). Can the system use intelligent decision making about how to process details in different ways in different places?
(One of many kinds of focusing of attention.)
- layered interpretations (image, low level image structures, silhouettes (2-d), parsed silhouettes, 3-d structures, affordances, many context-dependent features and relations.
- Attention is different in different parts of the system: different things are selected – **image features, objects, object features, locations in the image, locations in the world (relative to: room, table, object, object part...), relations, actions, ways of doing things, routes, other agents, social interactions,**

Further work to be done

All this is just a high level beginning: work still to be done includes:

- Developing theories of how a visual system can deal with input such as the cup and saucer pictures and produce representations of appropriate 3-D structures for use in simulations, either when perceiving actual motion or when thinking about possible motions:
pushing, throwing, stroking, denting, rubbing, smoothing ... all involve processes with different relations between hand or fingers and parts of surfaces.
- Collecting many more examples of types of simulative competence and developmental sequences, to give us some idea of what the various capabilities are and how they change over time.
- Coming up with a theory of how such capabilities could be implemented in a computer-based robot: we still have a very long way to go.
- Coming up with a theory of how such capabilities could be implemented in biological brains: we still have a very long way to go.
- Explaining what sort of innate meta-level capability is able to drive the processes of development in which these simulative capabilities, including new ontologies, new forms of representation, new modes of processing are acquired through interaction with the environment (exploration and play), along with help from older conspecifics.

31 Orthogonal competences (Added 1 Jan 2006)

By the time a typical child is about five years old, there is much detailed knowledge (some explicit, some implicit, and some a mixture) of several distinct kinds, which are orthogonal in the sense that their specific contents can vary independently, and which can be combined in different ways in creatively perceiving, understanding, and acting in the environment, i.e. perceiving and producing novelty.

This contrasts with animals (the majority of species I suspect) that can only learn associations involving **total** sensory arrays, or sub-patterns within such arrays (e.g. large blob getting bigger fast).

What follows is a first attempt to summarise some of the distinct kinds of competence involved in dealing with the environment, most of which have been illustrated in preceding sections.

A feature of the competences is that they can be combined in various ways that are creative insofar as they are novel to the individual. Although I emphasise the importance of this kind of creativity it should also be remembered that creatively developed competences can, through much practice, be 'compiled' into various sorts of habitual, or routine, specialised skills, perhaps represented in the same ways as the genetically determined, relatively inflexible, specialised skills of precocial species depending associations between global patterns.

It's not clear to me that 'orthogonal' is the best label for what I am talking about. Suggestions welcome.

Orthogonal environment-related competences 1

A typical child about five years old has much detailed knowledge of several distinct kinds, which can be combined in different ways in perceiving, understanding and planning actions in the environment:

- many **kinds of physical stuff** with different physical properties (e.g. water, sand, mud, wood, string, rope, paper, metal, stone, plastic, human skin, cotton wool, hair, butter, treacle, plastic film, aluminium foil, various kinds of food, wind, breath, fire and many more)
- different **kinds of surface features** – flat curved, smooth, rough, sticky, slimy, wet, sharp, textured in different ways, with ridges, furrows, dents, etc. etc.
This decomposes further into yet more orthogonal sub-spaces.
- different **shapes of whole objects**, varying in topological and metrical aspects, with both continuous and discrete sub-spaces, at different levels of abstraction,
E.g. there are discrete differences between numbers of holes, between being symmetric or not, having a long axis or not, etc. as well as a huge variety of types of continuous variation.
- different **ways in which new, possibly more complex, wholes can be formed by combining or modifying things** (in ways that depend on their shape, material, etc.)
We could include 'negative' combinations, e.g. gouging out, carving, punching a hole, to make a new shape as in sculpture.
Other shape-making transformations include bending, twisting, etc.

(continued...)

Orthogonal environment-related competences 2

.... Continued from previous page

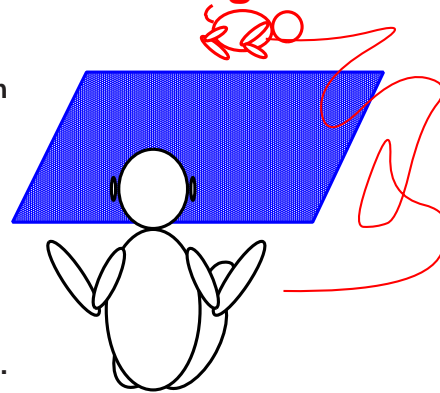
- different **sorts of spatial relations** between different objects of similar or different material (e.g. containing, touching, being glued to, being hooked round, being a certain distance apart, resting on, being mixed, attracting, repelling, etc. etc.)
There's a particularly important difference between 'rigid' containment (e.g. the streak of metal in a rock, the screw in a plank) and 'fluid' containment, e.g. water, sand or a small ball in a mug, a river flowing in its bed.
 - different **kinds of force** that can be applied to things, e.g. prodding, poking, stroking, squeezing, twisting, pulling, pushing, screwing, patting,
 - different **sorts of process** that can occur, including moving, rotating, changing shape, entering, coming out of, passing between, pushing, pulling, stretching, swaying, covering, uncovering, putting on (clothing), flocking, swarming, as well applying forces, and changing the application of forces
- Some of these may result from the individual's actions, some merely observed.
- As remarked previously, more complex things can be observed by an individual than produced by that individual, e.g. a busy street scene, a waterfall, a football match.
- (There may also be behaviours an animal (e.g. insect) can produce that it cannot perceive because its perceptual mechanisms lack the required sophistication.)

These lists are illustrative, not definitive or exhaustive, and do not include social abilities.

Example: Blanket and String

If a toy is beyond a blanket, but a string attached to the toy is close at hand, a very young child whose understanding of causation involving blanket-pulling is still Humean, may try pulling the blanket to get the toy.

At a later stage the child may either have extended the ontology used in its conditional probabilities, or learnt to simulate the process of moving X when X supports Y, and as a result does not try pulling the blanket to get the toy lying just beyond it, but uses the string.



However the ontology of strings is a bag of worms, even before knots turn up.

Pulling the end of a string connected to the toy towards you will not move the toy if the string is too long: it will merely straighten part of the string.

The child needs to learn the requirement to produce a straight portion of string between the toy and the place where the string is grasped, so that the fact that string is inextensible can be used to move its far end by moving its near end (by pulling, though not by pushing).

Try analysing the different strategies that the child may learn in order to cope with a long string, and the perceptual, ontological and representational requirements for learning them.

New Theory Vision Slide 85 Last revised: February 17, 2007 Page 117

Creativity in a physical environment

The different kinds of knowledge mentioned above can be combined in many different ways, including novel ways, in understanding what is perceived in the environment and what actions are and are not possible in different circumstances, and what the consequences of those actions will be.

We need to understand architectures and mechanisms for combining such knowledge and competences where appropriate.

Chapter 6 of *The Computer Revolution in Philosophy* attempted to analyse some of the processes about 30 years ago, but only at a high level of abstraction. <http://www.cs.bham.ac.uk/research/cogaff/crp/chap6.html>

- Sometimes competences are combined in **physical action**, using new combinations of material, tool, arrangement of parts or actions, to solve a problem; but in some cases it is done in thought (i.e. using deliberative mechanisms), as pointed out by Craik, Popper and many others.
- Precocial species, e.g. spiders, may have very specific 'hard wired' combinations of competence regarding specific kinds of stuff, specific spatial structures and processes; whereas humans some other altricial species are able both to **extend** knowledge within each of the categories, and to **forge new combinations** in perceiving novel scenes and performing novel actions — a meta-competence that underlies engineering, science and art.
- Such competence in pre-linguistic children and non-linguistic animals cannot depend on language, though it may be part of the basis for language, which, with other forms of cultural information-transmission (e.g. toys) enormously enhances and accelerates development.
- In a young child and in many animals the creative recombination of competence is applied in perceiving and using affordances for oneself, whereas humans later learn to see 'vicarious affordances', as discussed previously – essential in parents and carers watching children who may be about to hurt themselves, or may need help, or in seeing opportunities for predators who may attack one's young.

New Theory Vision Slide 86 Last revised: February 17, 2007 Page 118

As if this were not complex enough

Humans, though not infants, can combine all that creativity about the physical with creatively deployed knowledge about the mental. How?

- All of the 'orthogonal' competences listed above involve semantic competence: the ability to acquire, store, manipulate and use information.
- In humans at least there is also second-order and higher-order semantic competence (meta-semantic and meta-meta-semantic competence), namely the ability to use information about information and information users, e.g. thinking about what another individual, or oneself, can see, knows, wants, intends to do, about their reasoning, learning, planning or decision-making processes, including thinking about what A knows or fears about what B thinks about C.
- Such higher-order semantic competence is crucial to teaching.
- In teaching, learning from teachers, interpreting actions of others, negotiating, cooperating, planning revenge, and other social actions the opportunities for creative combination of previously listed competences with meta-semantic and social competences are enormous.
- E.g. understanding perceived unfamiliar behaviour in another may require combining knowledge (or hypotheses) about what the individual can and cannot see, what he can and cannot do, what he may wish to do, what affordances various physical materials shaped in a specific way can provide, and so on.
(**'Perhaps he dropped something through that grating and is trying to retrieve it using chewing gum on the end of a twig.'**)

How much of this applies to other animals?

- Not all animals can learn these things, even if they share a lot of physical structure with humans.
- So it is likely that there are very specific, very powerful brain mechanisms involved, possibly several different mechanisms that evolved in different combinations — we are not discussing all-or-nothing capabilities.
- Even among humans there may be different combinations, e.g. Archimedes, Shakespeare, Newton, Kant, Mozart, Darwin, Turing. Picasso, Menuhin – in which case there is no such thing as **human psychology**.
- If the hundreds, or thousands, of different kinds of knowledge acquired in the first few years are stored in different parts of the brain, using different mechanisms, then different sorts of brain damage or deficiency could interfere with different sub-competences. Has anyone looked? (**E.g. Williams' Syndrome?**)
- Since most of the creative brain mechanisms evolved before human language capabilities and appear in pre-linguistic children, despite involving rich forms of semantic and syntactic competence (using internal representations), it could be that the generative (combinatorial) and extendable aspects of those pre-linguistic competences provided a foundation for the later evolution of linguistic competence.

Perhaps that is an example of the common pattern in evolution: duplication of structures or mechanisms followed by differentiation. (See the 'primacy' paper.)

The relative unimportance of categories

There is a wide-spread tendency in many disciplines to think of mental processes, including perception, reasoning and action, as all involving the use of *categories* of objects.

This is natural if you think it is all done using propositions.

But from our viewpoint, which assumes that much of mental life involves running simulations at different levels of abstraction using the sorts of knowledge and competences that can be combined in different ways to generate diverse simulations, categorisation of things is not so important: what generates and constrains many kinds of processes is kinds of matter, structures, relationships and modes of composition.

Multi-strand relationships and multi-strand processes are specially important.

Categorisation of types of objects, situations, actions, etc. may come later as the need to add certain ways of 'chunking' reality arises.

Do insects categorise things?

Perhaps we are all insect-like in more ways than we suspect.

An old idea: progressive deepening

The tasks are VERY difficult.

To explore integration of different components it may be necessary to simplify some of the components temporarily.

Compare Joe Bates on "broad architectures".

<http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/papers/sigart.2.4.ps>

Progressive deepening can follow.

Applications for a child-like robot.

A good working model of generic child-like intelligence, including early forms of self-understanding, could lead to important new explanations of both earlier and later stages of development, and could be the basis of many different sorts of demanding practical applications.

Example applications include

**guide-robots for the blind,
more flexible assembly robots,
entertainment robots (and synthetic agents in virtual worlds),
search and rescue robots,**

**home assistants for elderly or disabled people
who don't want to be a burden on other people
(a growing subset of the population).**

All this is part of the motivation for the CoSy project.

33 We need to think about architectures

The sort of system we are discussing has many components doing many different things in parallel. Putting the pieces together in working architecture is a non-trivial task for engineers and for scientists attempting to produce explanatory models.

We need good theories about the space of possible architectures, and good theories about particular architectures in that space in order to explain the wide variety of biological phenomena and in order to understand the development of humans, since we are not born with a fully fledged architecture: they architecture grows in ways that may partly replicate some of our evolutionary history but will be much influenced by our culture and physical environment.

In other documents available at Birmingham I have (with the help of colleagues) developed this idea in much more detail, including showing, for example, how different aspects of motivation and emotion relate to different architectural layers with different competences.

Architectural challenges

One requirement for progress is specification of a virtual machine architecture that can combine many known kinds of human capabilities, including

- evolutionarily very old **reactive** mechanisms, also used for highly trained, semi-automatic learnt competences derived from results of deliberative and reflective processes (after practice)
- newer **deliberative** mechanisms and
- biologically rare **reflective, meta-management** mechanisms with meta-semantic capabilities (the ability to represent processes in things that themselves represent other things, unlike rocks, trees, levers, wheels, blocks, ...).

Papers and presentations in the Cognition and Affect project provide more detailed analyses of these architectural features, illustrated on the next slide. See

<http://www.cs.bham.ac.uk/research/cogaff/>
<http://www.cs.bham.ac.uk/research/cogaff/talks/>
<http://www.cs.bham.ac.uk/research/projects/cosy/papers/>

It seems that different sorts of simulations may occur in different parts of the architecture. E.g. it may be the reactive component that makes people 'kick' while watching an exciting football match.

The three crude categories can be further sub-divided as shown in Minsky's more fine-grained architectural sub-divisions in 'The Emotion Machine'.

A hypothetical Human-like architecture:

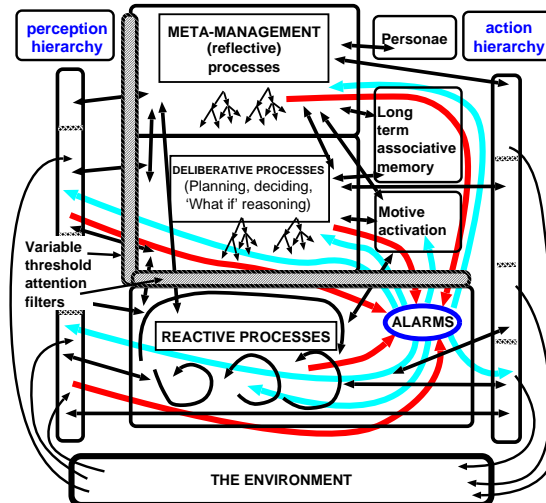
H-CogAff (See <http://www.cs.bham.ac.uk/research/cogaff/>)

This is an instance (or specialised sub-class) of the architectures covered by a generic schema called "CogAff". Many required sub-systems are not shown.

Where could it come from?

Various trajectories:

- evolutionary,
- developmental,
Altricial species build their architectures while interacting with the environment?
- adaptive,
- skills developed through repetition (how?)
- social learning, including changing personae...



(This is an illustration of some recent work on how to combine things: much work remains to be done. This partly overlaps with Minsky's *Emotion machine* architecture.)

For more details, see the presentations on architectures here

<http://www.cs.bham.ac.uk/research/cogaff/talks/>

But that's just one example

WE NEED LOTS MORE WORK ON A TAXONOMY OF TYPES OF ARCHITECTURE,

based on analysis of

- **Requirements** for architectures,
- **Designs** for architectures,
- **Components** of architectures
 - Varieties of information structures
 - Varieties of mechanisms
 - Kinds of control systems
- Ways of assembling components
- How architectures can develop,
- Tools for exploring and experimenting with architectures

There is something deep and important about 3-D spatial perception and understanding

CONJECTURE:

Several different aspects of our ability to perceive and manipulate structured 3-D objects have, during biological evolution, profoundly impacted on the forms of representation available to us for a variety of tasks (including non-spatial tasks), the ontologies we cope with, the architectures used in human and some other animal minds, and our understanding of causation.

Some of this is shared with other animals, including primates, hunting mammals, and some nest-building birds.

Explaining how this works is a pre-requisite for developing useful human-like domestic robots (though that is not my main goal).

Request for help

Please join this project, if you can.

It is very difficult and requires contributions from many deep, creative thinkers (including designers), from many disciplines.

More things to show

Pythagoras

(At Birmingham run the program (inspired by Norman foo, who has his own version) showing the Chinese proof of Pythagoras theorem on a Sun or linux machine:

`/home/staff/axs/temp/pythag`

The program requires Poplog, which is available to anyone with a PC running linux, or Solaris on a Sun.)

Sheep

(Demonstrate a program where causal relations can be changed.)

Child and tunnel

(Show video of child playing with a toy train that goes through a tunnel)