Expanded version of presentation at
BBSRC Computational Modelling Workshop – Birmingham – May 31st–June 2nd, 2007
Closing the gap between neurophysiology and behaviour: A computational modelling approach
http://comp-psych.bham.ac.uk/workshop.htm

Last updated August 31, 2007

Architectural and representational requirements for seeing processes and affordances.

Aaron Sloman

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

With help from Jackie Chappell and colleagues on the CoSy project

These slides are accessible from here:
http://www.cs.bham.ac.uk/research/cogaff/talks/#compmod07
Along with other related presentations and papers here:
http://www.cs.bham.ac.uk/research/projects/cosy/papers/

WARNING: My slides have too much detail for presentations. They are intended to make sense if read online.
NOTE: These slides are produced using LaTex (on Linux), not powerpoint.

Expanded Abstract (1)

Over several decades I have been trying to understand requirements for a robot to have human-like, or more generally, animal-like, visual competences. I have written several papers pointing out what some of those requirements are and how far all working models known to me are from satisfying them. Many computational, neural and psychological theories address only a small subset of those requirements.

For example, human visual competence is used in imagining things that differ from what is seen, including imagining processes that change spatial structures. Doing elementary Euclidean geometry involves seeing how processes of construction can produce new geometric structures from old ones in proving theorems, such as pythagoras' theorem. Understanding how an old-fashioned clock works usually involves *seeing* causal connections and constraints related to possible processes that can occur in the mechanism. Seeing why turning a nut makes it move along a screw requires understanding relationships between one curved 3-D surface sliding against another. These and other examples use our ability to see how causal influences propagate *spatially* through a mechanism. Such visual reasoning is very different from logical reasoning, but is a form of reasoning (Sloman 1971, 1978(ch7), 2005b).

The more general point is that vision has multiple functions that are not always listed in textbooks on vision (Sloman 1982, 1989, 2001), including

- various kinds of *control* functions (e.g. posture control, triggering saccades, triggering alarms or sexual responses, or aesthetic responses, sight-reading music, continuous visual servoing while reversing a car, grasping an object, tracing a drawing, etc.) in addition to
- *descriptive* functions, e.g. answering a question, locating an object, detecting a threat, providing causal explanations, helping speech comprehension, helping with predicting and planning, checking conditions during plan execution, and checking predictions

(These are often mis-described as 'where vs what' functions, 'how' vs 'what', or 'action vs perception' functions.)

Content perceived in descriptive functions
It has been evident for many decades that percepts can involve hierarchical structures. Many of those are representable as loop-free trees, at a coarse level, E.g. most animals and plants are decomposable into parts with appended parts that have appended parts, etc. However many artifacts, e.g. bicycles, do not fit that model, and neither do internal structures of plants and animals. So representation of what is perceived can include cyclic graph structures as well as tree structures. However more complex 'multi-strand' relationships can hold between a group of objects or parts of an object with different parts related in different ways.

Less obviously, visual and other forms of perception can involve layered ontologies. E.g. one sub-ontology might consist entirely of 2-D image structures and processes, whereas another includes 3-D spatial structures and processes, and another includes kinds of 'stuff' of which objects are made (e.g. mud, wood, skin, meat, banana, paper,...) and their properties (e.g. rigidity, elasticity, solubility, temperature, thermal conductivity, etc.). A more abstract ontology is required for seeing social and mental states and processes, e.g. seeing people fighting, or exchanging objects or collaborating on some task, or seeing an individual as happy or sad, or as intently watching a small moving object. The use of multiple ontologies is even more obvious when what is seen is text, or sheet music, perceived using different geometric, syntactic, and semantic ontologies.

Continued ...

Abstract (2)

In 2005 when I was working on an EU-funded robot project (CoSy) I noticed what should have been obvious long before, namely that

(a) the content of what is seen is often processes and process-related affordances – i.e. possibilities for and constraints on processes (as illustrated later), and

(b) the content of what is seen usually involves both hierarchical structure and multiple ontologies.

Together (a) and (b) imply that vision often involves perception of processes, and possibilities for and constraints on processes, in which changes occur concurrently at different ontological levels.

These slides expand the above points by presenting examples of requirements for artificial visual systems with the same functions as human vision. The requirements also determine criteria for evaluating theories purporting to explain how human vision works.

No claim is made about completeness, however, or precision!

One way to make progress may be to start by relating human vision to its many evolutionary precursors, including vision in other animals. If newer systems did not replace older ones, but built on them, we should look for different kinds of visual processing going on concurrently, especially when a process is perceived that involves different levels of abstraction perceived concurrently, e.g. continuous physical and geometric changes relating parts of visible surfaces and spaces at the lowest level, discrete changes, including topological and causal changes at a higher level, and in some cases intentional actions, successes, failures, near misses, etc. at a still more abstract level.

The different levels use different ontologies, different forms of representation, and probably different mechanism, yet they are all interconnected, and all in partial registration with the optic array (though not with retinal images, e.g. because perceived processes survive saccades, which rules out use of retinotopic maps for the representation of such processes

[See A.Trehub The Cognitive Brain (MIT Press, 1991) now online at `http://www.people.umass.edu/trehub/`]).

Moreover, those descriptive functions can be performed concurrently with control functions, including visual servoing, posture control, triggering saccades, etc., for example when you bake a cake while describing things as they happen.

It may turn out that models designed to explain small subsets of these phenomena that are strongly constrained by experimental conditions are incapable of being expanded to explanations of the whole range of functions.

The problem

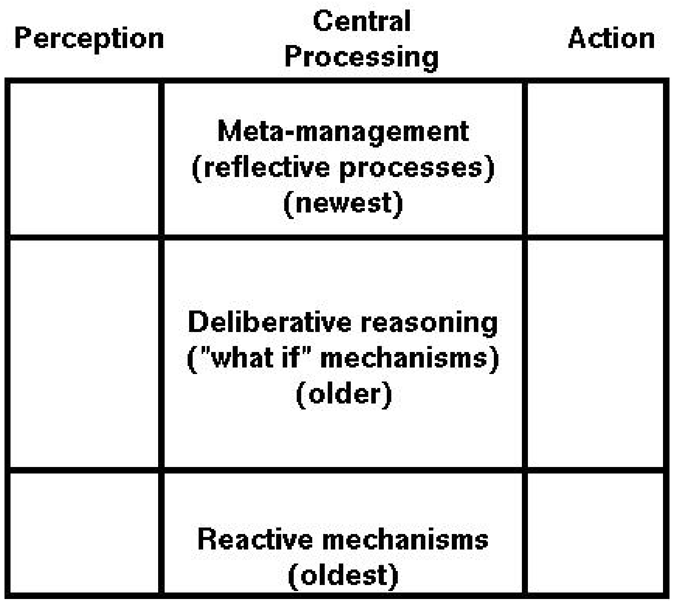
- Despite the aims of the workshop I shall not say much about what neural mechanisms are or how they work (except for a brief speculation, later): I am mainly concerned with what sort of virtual machine is implemented on them – especially what its visual functions are.
- Human researchers have only very recently begun to understand the variety of possible information processing systems.
- In contrast, for millions of years longer than we have been thinking about the problem, evolution has been exploring myriad designs.
- Those designs vary enormously both in their functionality and also in the mechanisms used to achieve that functionality

  including, in some cases, their ability to monitor and influence some of their own information processing (not all).
- Most people investigating natural information processing systems assume that we know more or less what they do, and the problem is to explain how they do it.
- But perhaps we know only a very restricted subset of what they do, and the main initial problem is to identify exactly what needs to be explained.
- A piecemeal approach may lead to false explanations: working models of partial functionality (especially functionality in artificial experimental situations) may be incapable of being extended to explain the rest.

The CogAff Schema (for designs or requirements)

Requirements for subsystems can refer to

- **Types of information handled:** (ontology used: processes, events, objects, relations, causes, functions, affordances, meta-semantic states, etc.)
- **Forms of representation:** (transient, persistent, continuous, discrete, Fregean (e.g. logical), spatial, diagrammatic, distributed, dynamical, compiled, interpreted...)
- **Uses of information:** (controlling, modulating, describing, planning, predicting, explaining, executing, teaching, questioning, instructing, communicating...)
- **Types of mechanism:** (many examples have already been explored – there may be lots more ...).
- **Ways of putting things together:** in an architecture or sub-architecture, dynamically, statically, with different forms of communication between sub-systems, and different modes of composition of information (e.g. vectors, graphs, logic, maps, models, ...)



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

In different organisms or machines, the ‘boxes’ contain different mechanisms, with different ontologies, functions and connectivity, with or without various forms of learning.
In some the architecture grows itself after birth.

In microbes, insects, etc., all information processing is linked to sensing and acting, and all or most information about the current environment is only in transient states in reactive mechanisms, whereas for more sophisticated organisms, evolution discovered the massive combinatorial advantages of exosomatic, amodally represented, ontologies, allowing external, future, past, and hypothetical processes, events and causal relations to be represented.

Perhaps “mirror” neurones – should be called “exosomatic abstraction” neurons?

Can we use brain structure as a guide to architecture?

- Some people may be tempted to assume that any accurate architecture must reflect brain structure.
- That could tempt them to assume that an architecture diagram should be labelled with known portions of brains.
- There are two problems with this:
  - it does not allow us to specify a virtual machine information-processing architecture that is common to an animal with a brain and a machine that uses artificial computational mechanisms.
  - it does not allow for the possibility that high level functions don't map onto separable parts of brains but are implemented in a more abstract way (just as data-structures in a software system may not map onto fixed parts of a computer's physical memory, e.g. if virtual memory and garbage collection mechanisms are used.

Anyhow the attempt to specify an architecture that I talk about makes no assumptions about how the components map onto brain mechanisms.

Rather it can be construed as a specification of a large collection of requirements for something to function as a certain kind of thing, e.g. an adult human, an infant human, a nest-building bird, or whatever we are trying to explain.

Of course similar, or partly similar, architectures may be useful in some kinds of intelligent robots, or plant-control systems, or ....

A special case of the CogAff schema

Since about 1978 I have been trying to characterise varieties of functionality required for a human-like architecture. The H-CogAff diagram illustrates some of the results, at a high level of abstraction.

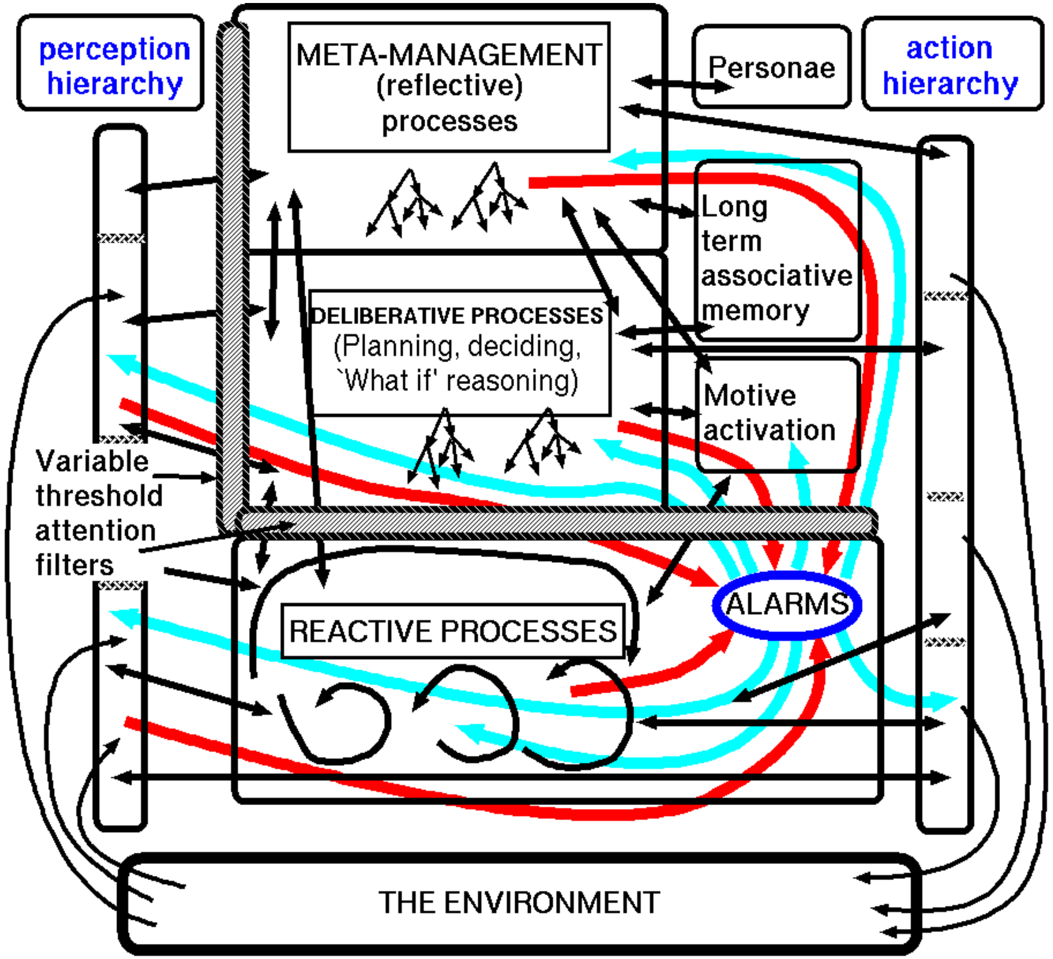
Regard this as an architecture for a collection of requirements.
We can use this to derive different architectures for different organisms/robots, depending on which requirements are important: a space of possibilities.

There are partial implementations of designs meeting different subsets of these requirements, using our SimAgent toolkit.

The architecture, and the more general CogAff scheme are described in more detail in many papers and presentations in the Birmingham Cogaff web site.

This overlaps a lot with Minsky's Emotion Machine architecture but we use different principles of subdivision.

More information is available


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

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

http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html

That’s just one example

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

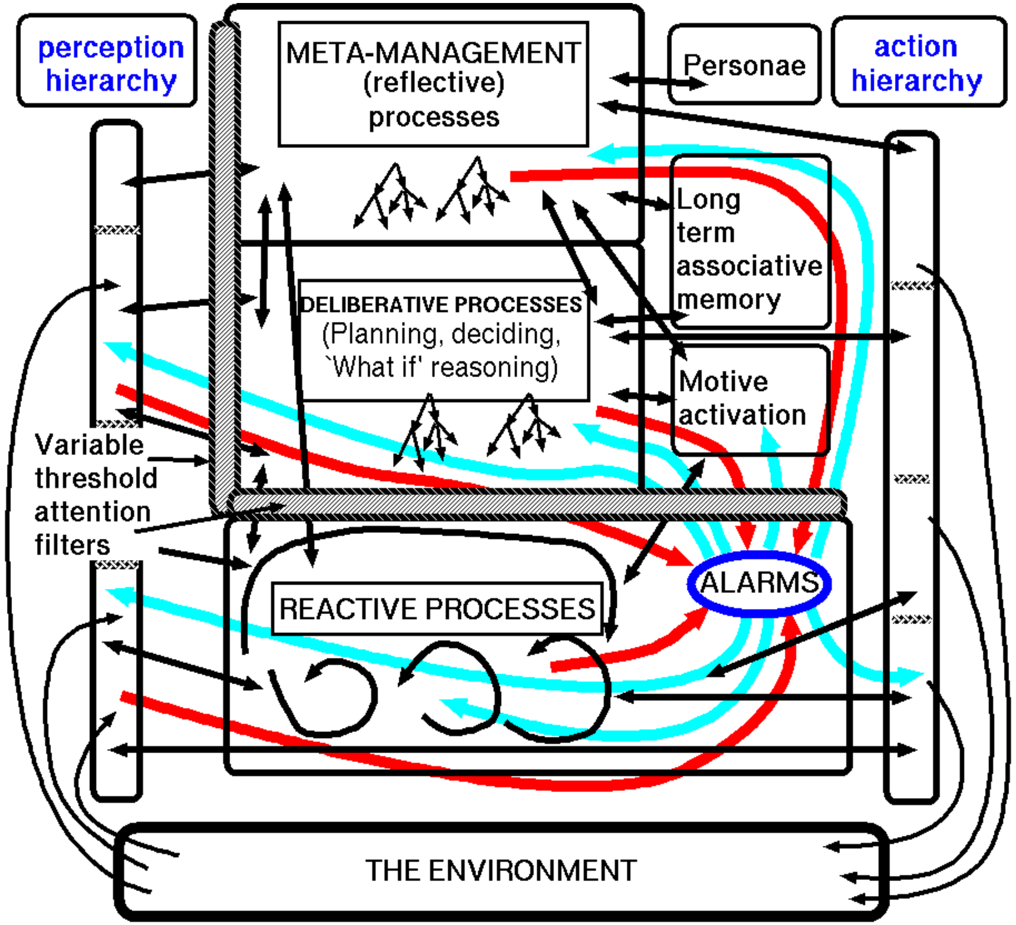
based on analysis of

- Requirements for architectures,
- Designs for architectures,
- Components of architectures
  - Varieties of information structure
  - Varieties of mechanisms
  - Kinds of control systems
  - Ontologies and forms of representation needed in different subsystems
- Ways of assembling components
- How architectures can develop,
- Tools for exploring and experimenting with architectures
- We also need agreed diagrammatic conventions.

More about H(human)-CogAff

H-CogAff specifies a sub-class of the architectures covered by the “CogAff” schema.

This is a sketchy indication of some of the required subsystems and how they are connected.
It is hypothesised that both vision and action subsystems have different (concurrently active) layers of functionality related to the different central layers/mechanisms, concerned with different tasks, ontologies, and forms of representation.


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

An architecture can have complex origins with different trajectories:

- evolutionary,
  precocial species get it all from the genome
- developmental,
  altricial species build their architectures, ontologies, knowledge, etc. while interacting with the environment
- adaptive parameter adjustments,
  many forms of statistics-based learning
- skills compiled through repetition
  learning to grasp, walk, read music, play tennis,...
- social learning, including changing personae...

Much work remains to be done. See the presentations on architectures here
`http://www.cs.bham.ac.uk/research/cogaff/talks/`

The role of visual mechanisms in the architecture

The rest of this presentation focuses on aspects of the architecture and the capabilities involved in the architecture that relate to human vision.

Visual and spatial cognition

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).

CONJECTURE

Mechanisms for perception of the 3-D environment penetrate deep into the cognitive system, and cognitive mechanisms penetrate deep into the perceptual subsystems.

Views on functions of vision

- 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)

    .... and more ....

- I'll try to present phenomena that require a richer deeper theory.

  It will be evident that an adequate theory must use many of the above ideas, and assemble them in new ways with some new details.
  Some of the old ideas can be criticised.
  There are important implications: 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.

Some themes in what follows

- Hierarchically organised processes as well as structures are seen
- Ontological layers are involved, in addition to taxonomies and part-whole hierarchies
- There are different sorts of seeing, with different perceptual contents:

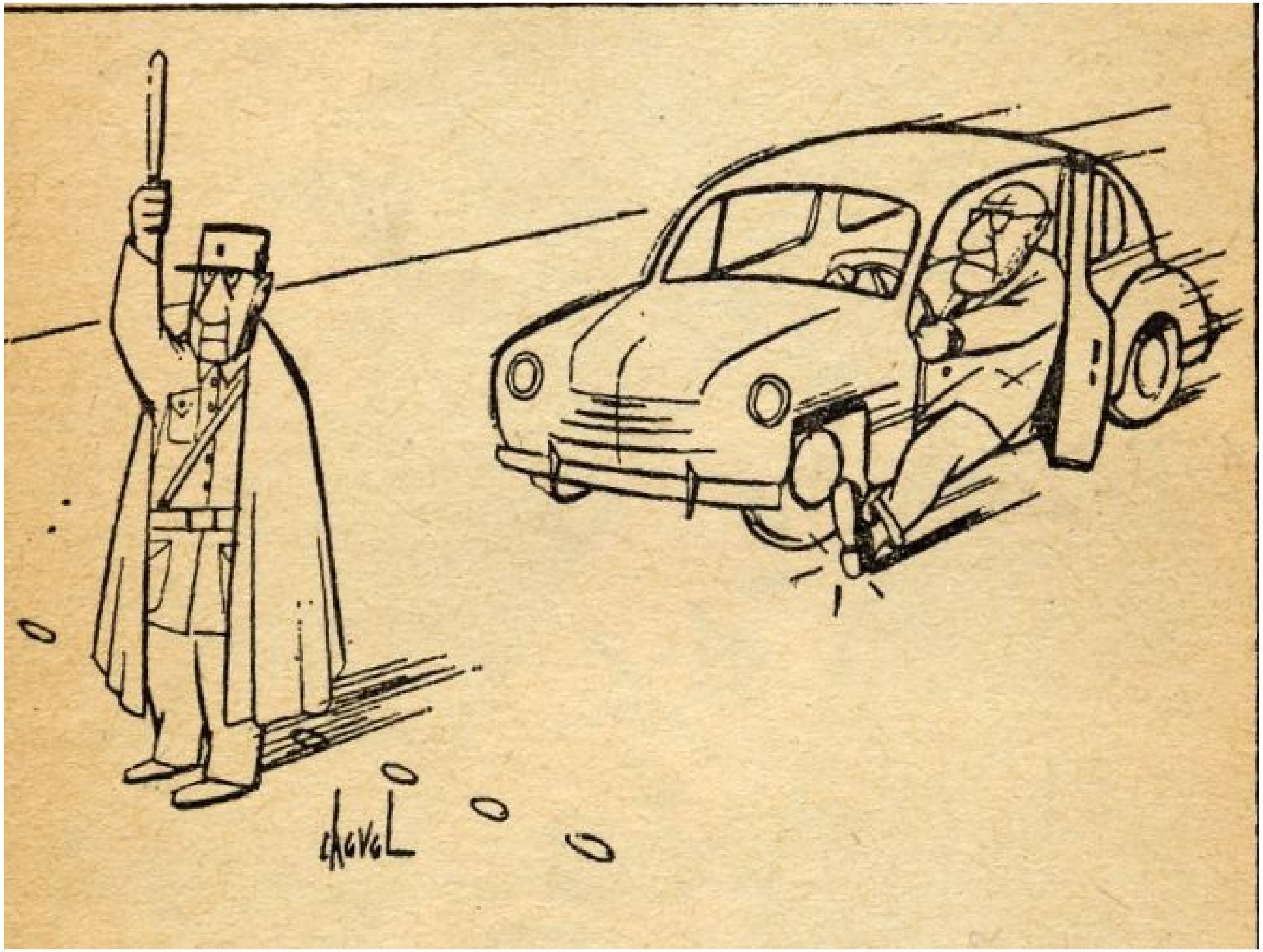
  geometrical structure, kinds of stuff, what's happening, causal interactions, mental states, musical or mathematical meaning.
- Amodal, exosomatic, representations are needed:

  e.g. for expert and generalisable grasping, and manipulating
- Multi-strand relationships are seen: continuous, discrete, logical
- Multi-strand multi-level processes are seen: continuous, discrete, logical

  Are visual systems primarily concerned with perception of processes?
- Do we really understand the full variety of functions of vision?
- Ontologies and forms of representation
- Labyrinthine, not modular architectures are needed (1989 paper)
- We need to be comparative – learn by comparing and contrasting:

  humans and other species
  humans at different stages
  humans with different pathologies
  humans and many kinds of possible machine

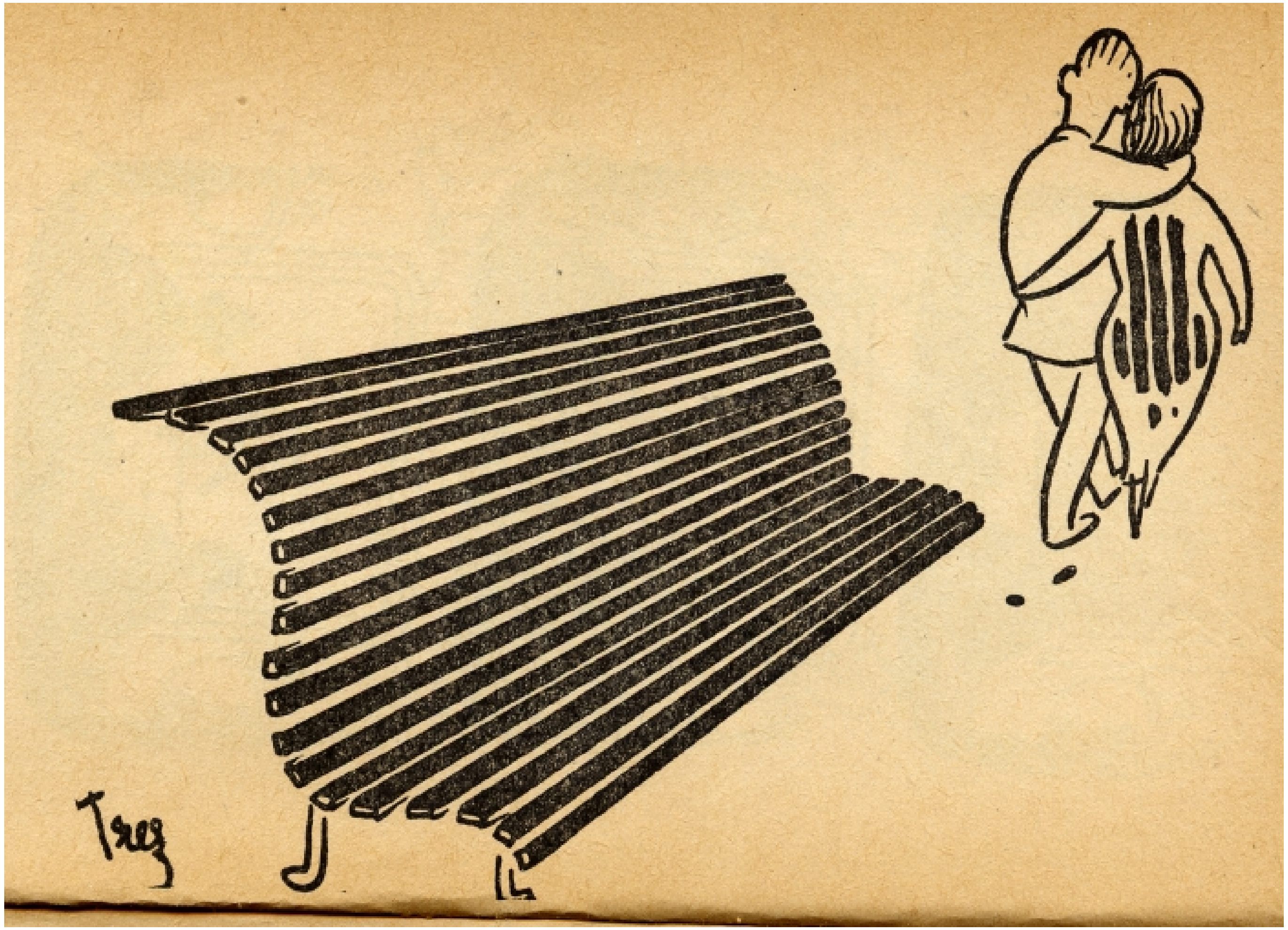
What do you see?



Perhaps you see a process extending to a future time?

And causal connections?

What do you see?

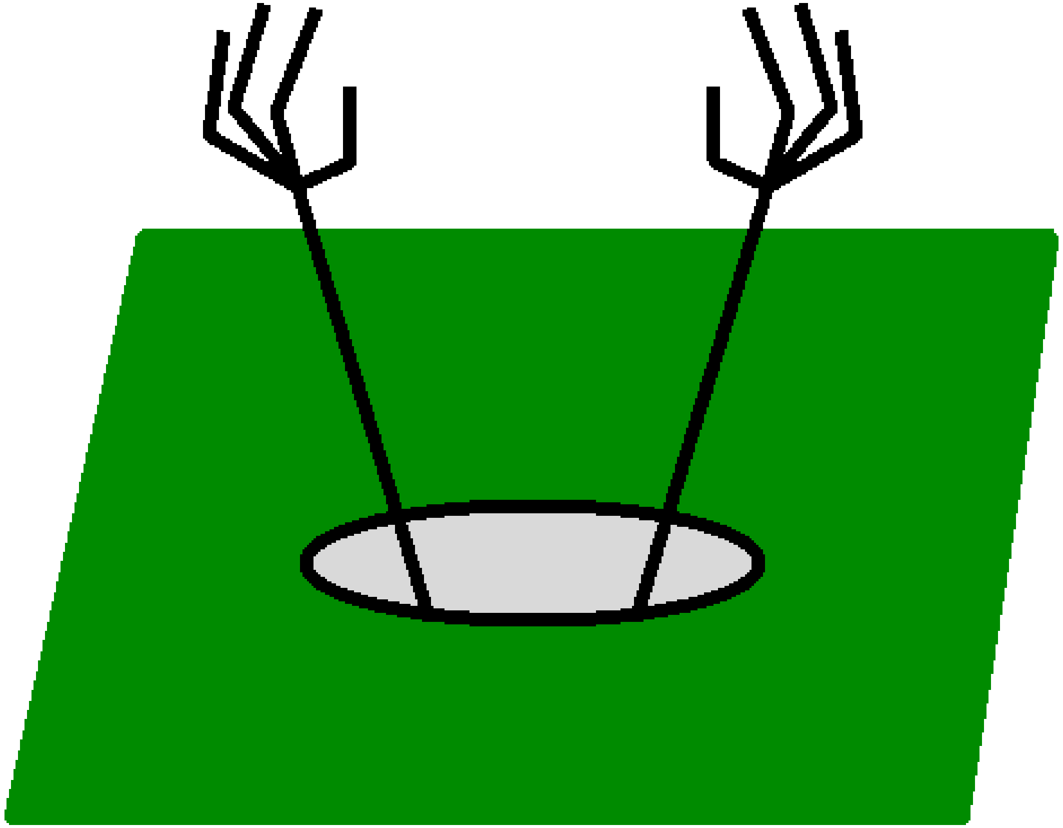


Various objects and relationships of different sorts

Perhaps you see a process starting at an earlier time?

And causal connections?

What do you see?



In many cases what you see is driven by the sensory data interacting with vast amounts of information about sorts of things that can exist in the world.

But droodles demonstrate that in some cases where sensory data do not suffice, a verbal hint can cause almost instantaneous reconstruction of the percept, using contents from an appropriate ontology, including causes and effects.

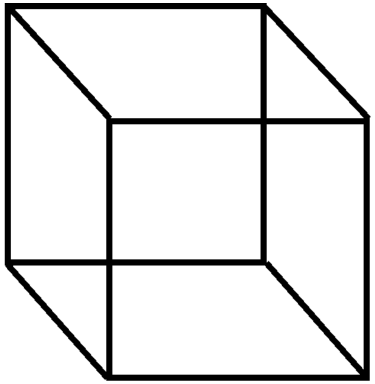
See also http://www.droodles.com/archive.html

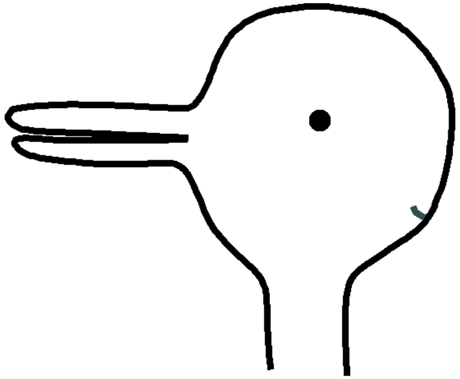
Possible captions for the figure: ‘The early worm catches the bird’, or ‘Early bird catches very strong worm’.

Seeing beyond the retina

What we see is not all derived from retinal images, as shown by ambiguous images:

what is seen flips even though what is on the retina does not change.





Some things not in a retinal image are described as seen, not inferred: WHY?

Examples: relative depths, and 3-D orientations of parts of the cube, parts of an animal, which way an animal is looking, ...

An ontology is involved in every percept – usually several ontologies.

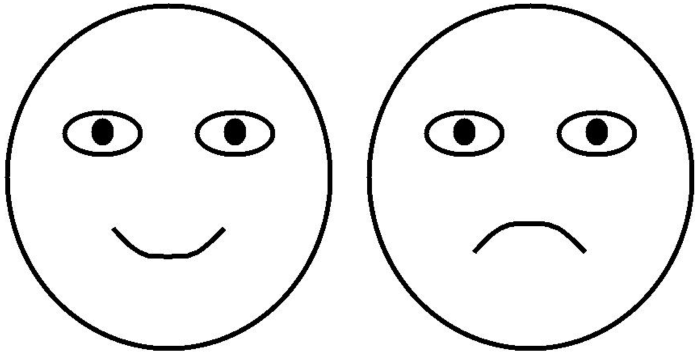
Different dimensionalities and ontological layers

Some ontologies are 2-D only – e.g. line, junction, ...

Some involve 3-D structures and relations that can change while the contents of the optic array do not,

e.g. relative distance, 3-D orientation of lines (sloping up and away *vs* down and away), etc.

What sort of ontology is needed to describe the flips in the duck-rabbit?
Geometry is not enough.



Some percepts use a meta-semantic or mentalist ontology: (including 'happy', 'sad', 'looking left', 'looking right'.)

Which way an animal is looking is related to what information it can get from the environment. (Important for prey and predators.)
We can also see causation (discussed below).

Perception vs inference

In perception the appropriate ontology is deployed to construct an interpretation whose details are in partial registration with the sensory array or some systematic transform of it.

(Not a retinal image, but the optic array.)

This implies that in the case of vision the perceptual contents are closely related to point of view and direction of view.

What is seen changes systematically, globally, and in an intricate way, with movements of the head or eyes.

This is probably the main fact that led to the sensorimotor theory of consciousness (e.g. O'Regan and Noë).

However the relationships between sensory contents and location are not quite as intimate for touch, hearing, smell, and feeling of ambient temperature.

Moreover even in the case of vision, insofar as what is seen includes the structure and properties of objects those require an exosomatic (non-sensorimotor) ontology.

There are specialised forms of representation for combining spatial topological and causal properties and relationships in both static structures and in multi-strand processes, that we do not yet understand.

Except that their properties are very different from Fregean, logical representations.

(In 1971 I called them 'analogical' representations'.)

Moreover, the process uses dedicated, specialised mechanisms operating on the particular sensory input and the interpretations, as opposed to general purpose inference mechanisms.

However there may be some fuzziness in the perception/inference boundary.

Example: Multiple perceptual routes

H-CogAff specifies

multi-window perception (different levels of perception occurring simultaneously)

and

multi-window action, (different levels of action occurring simultaneously)

whereas many architectures assume peephole perception and action.

In humans, and probably some other species, the visual and action sub-systems have architectural layers (evolved or developed) that handle ontologies at different levels of abstraction (including in some cases percepts and actions related to mental states of oneself and others).

To support this, visual and action subsystems need to have multiple connections to different sorts of central sub-systems, as well as to other sensory and motor subsystems.

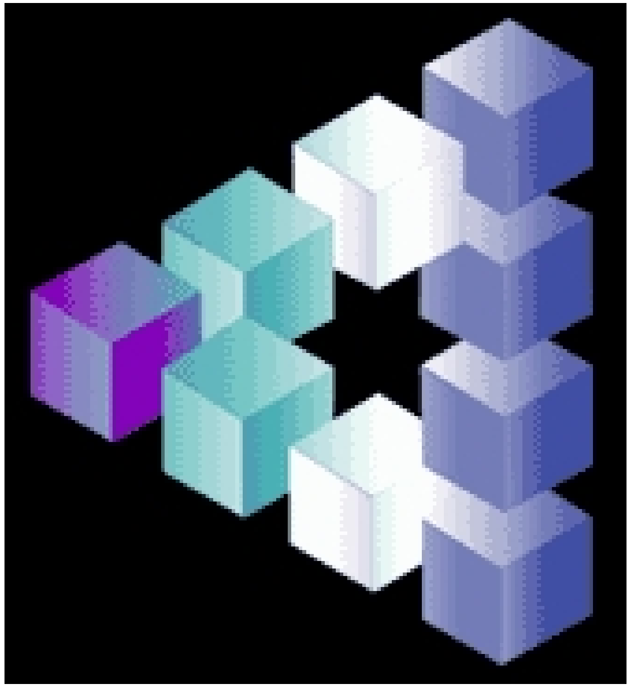
(As shown in the diagrams.)

So, instead of one or two (e.g. dorsal and ventral) routes from vision, we have multiple routes,

e.g. to blinking reflexes, saccade generators, posture control subsystems, visual servoing mechanisms, question answering mechanisms, planning mechanisms, prediction mechanisms, explanation constructors, plan execution mechanisms, learning mechanisms (in several different architectural layers), alarm subsystems, communication mechanisms, social mechanisms.

Similar comments apply to connections with action sub-systems.

High level percepts can be inconsistent



High level percepts can be inconsistent

(Picture by Reutersvard, in 1934 – before Penrose)

This tells us important things about the visual system – and some of the contents of visual consciousness.

What you see is not only what exists, but multiple affordances.
Think of all the things you can do with or between the little cubes.
If the picture were huge, you might never discover the impossibility

Collections of affordances can be inconsistent: but not models of a scene.

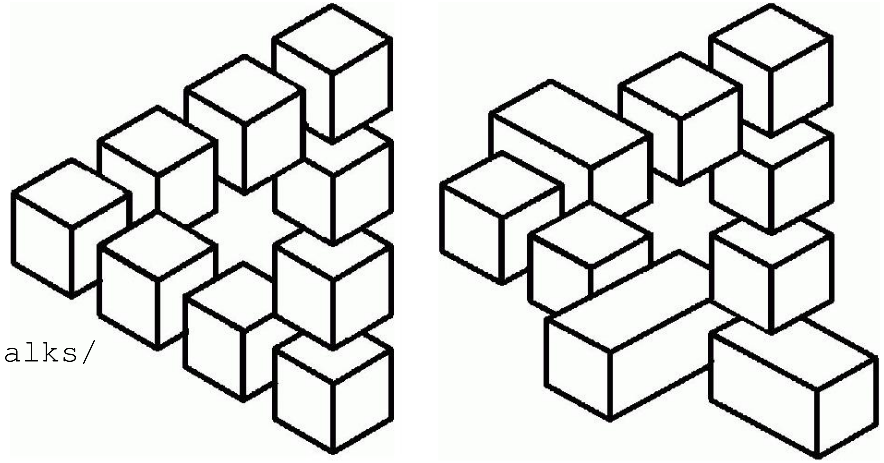
Compare Escher’s pictures, e.g. the Waterfall.

For more on visual processing see

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

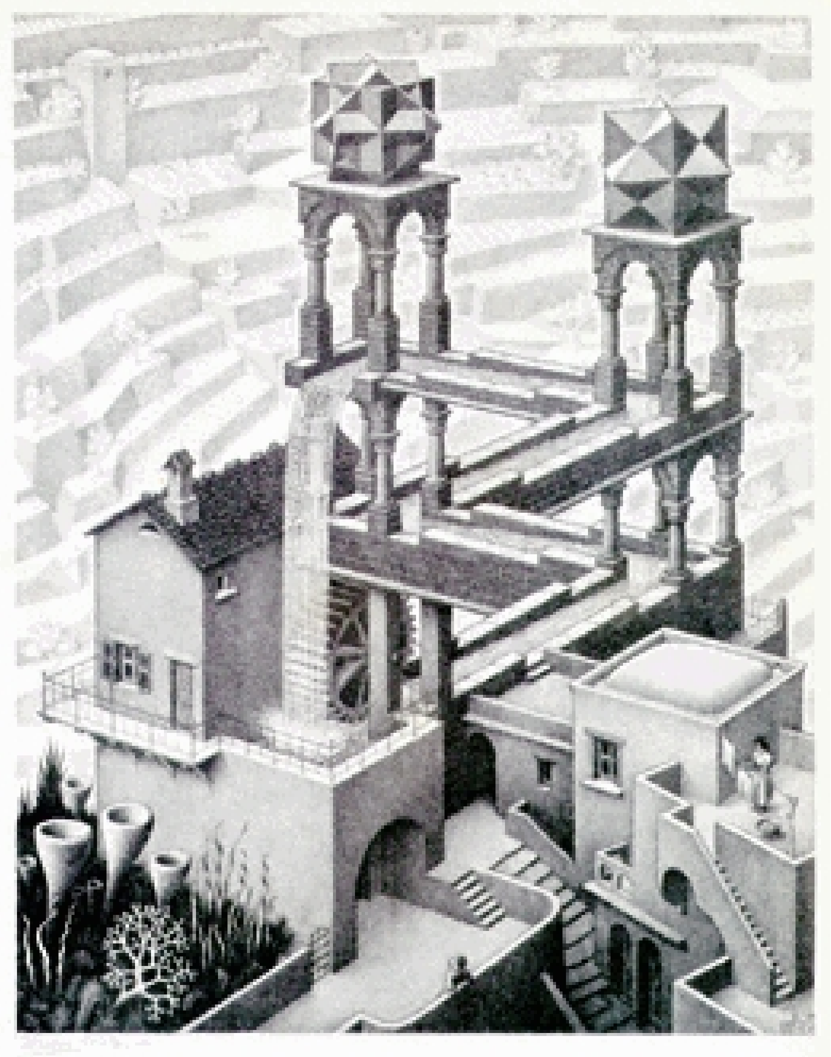
Why doesn’t a young child see the inconsistency?

What has to change to make the inconsistency evident?



If I gave you a box containing many cubes and rectangular blocks -- would you be able to build objects like these?

Escher's Weird World



Many people have seen this picture by M.C. Escher:

a work of art, a mathematical exercise and a probe into the human visual system.

You probably see a variety of 3-D structures of various shapes and sizes, some in the distance and some nearby, some familiar, like flights of steps and a water wheel, others strange, e.g. some things in the 'garden'.

There are many parts you can imagine grasping, climbing over, leaning against, walking along, picking up, pushing over, etc.: you see both structure and affordances in the scene.

Yet all those internally consistent and intelligible details add up to a multiply contradictory global whole. What we see could not possibly exist.

There are several 'Penrose triangles' for instance, and impossibly circulating water.

Can you see the contradictions? They are not immediately obvious.

A visual percept cannot be a model

Models cannot be inconsistent

However if percepts are made up of fragments combined in a manner that does not correspond to full spatial integration then inconsistencies are possible.

E.g.

- A is bigger than B
- B is bigger than C
- C is bigger than A

Why might his be desirable?

Because the very same scene needs to be perceivable in different ways, depending on current goals, interests, etc.

And each item of information needs to be capable of being combined with many others, for different purposes.

So it must be possible to switch different items of information in and out of the percept.

E.g. different affordances.

This flexibility is not possible if the percept is just a ‘holistic’ copy or a projection of all the structure of the scene.

In any case such a copy or projection could not explain perception because it would still need to be interpreted in order to be usable, e.g. for predicting, planning, control of action, etc.

Types of internal language

The information acquired through perceptual processes (and other means) needs to be represented somehow, in order to be usable.

The uses made by many animals, and by young children imply that the information acquired through perception is represented in mechanisms that can cope with

- Rich structural variability

  e.g. a nest built by a crow constantly changes its structure during construction.

- Compositional semantics

  Seeing new structures as made up of familiar kinds of components in varying relationships.

  The composition need not be logical: can be geometric or topological.

Thus prelinguistic children and some other animals must have internal languages with features normally associated with human communicative languages.

(Sloman 1979) `http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#43`

We need a generalised notion of language: a G-Language.

"G-language" defined in Sloman and Chappell BBS to appear (2007)

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

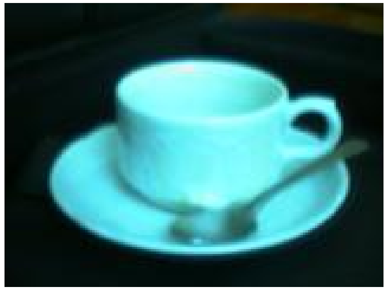
Conjecture:

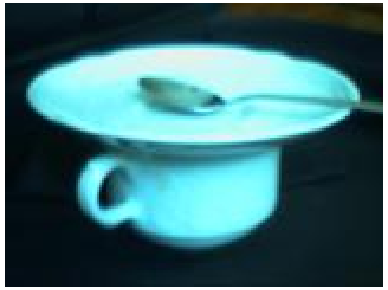
G-languages with rich structural variability and compositional semantics needed for internal use in perception, reasoning, planning, goal formation, etc., evolved before external languages for communication evolved. and also develop in young children before human languages are learned.

That's how children can have something they want to communicate, to drive the language learning process.

Some tasks for a crow-challenging robot?

We can begin to analyse some of the representational requirements for visual systems by analysing tasks for which vision can be used.

(a) 

(b) 

Using a two-finger gripper, what actions can get from (a) to (b) and back again?
Or with saucer upside down?

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
- identify possible actions: grasping this thing here from this direction

  (deliberative premeditation seems to use an action schema with approximate, qualitative parameters instead of definite actual parameters that would be used if the action were performed.)
- think about various effects of actions, including changing effects of continuous processes

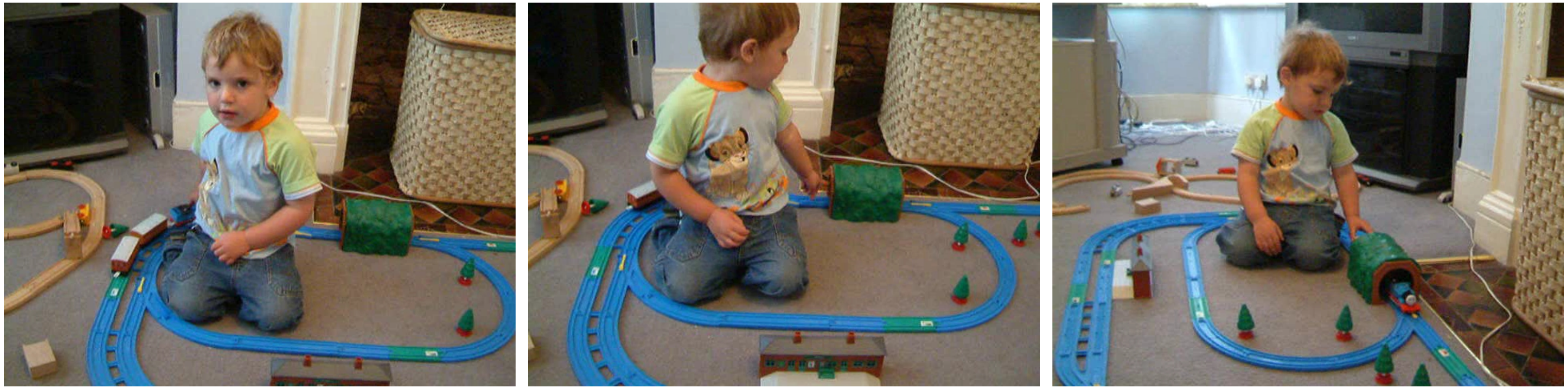
NOTE: there are problems here partly analogous to problems of reference and identification in language, except that the mode of reference is not linguistic and what is referred to typically cannot be expressed in language because it is anchored in non-shared structures and processes.

(Internal 'attention' processes are partly like external pointing processes: virtual fingers. (Pylyshyn 197??))

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 and comparing their merits.

Snapshots from tunnel video

A child playing with his train illustrates many unobvious functions of vision.



- 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?

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.
- Conjecture: our architecture evolved to support at least three sorts of concurrency:
  - Perceiving multiple concurrent external processes
  - Representing the same external process at different levels of abstraction
  - Different concurrent actions internal to the individual that use different parts of the information-processing architecture simultaneously, such as
    - controlling walking (including posture control),
    - working out where to walk,
    - performing thought processes required for discussing philosophy or the view or .... with a companion,

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 ‘hierarchical 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 different levels were not just levels in a part-whole hierarchy, but involved different spaces, different ontologies.

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

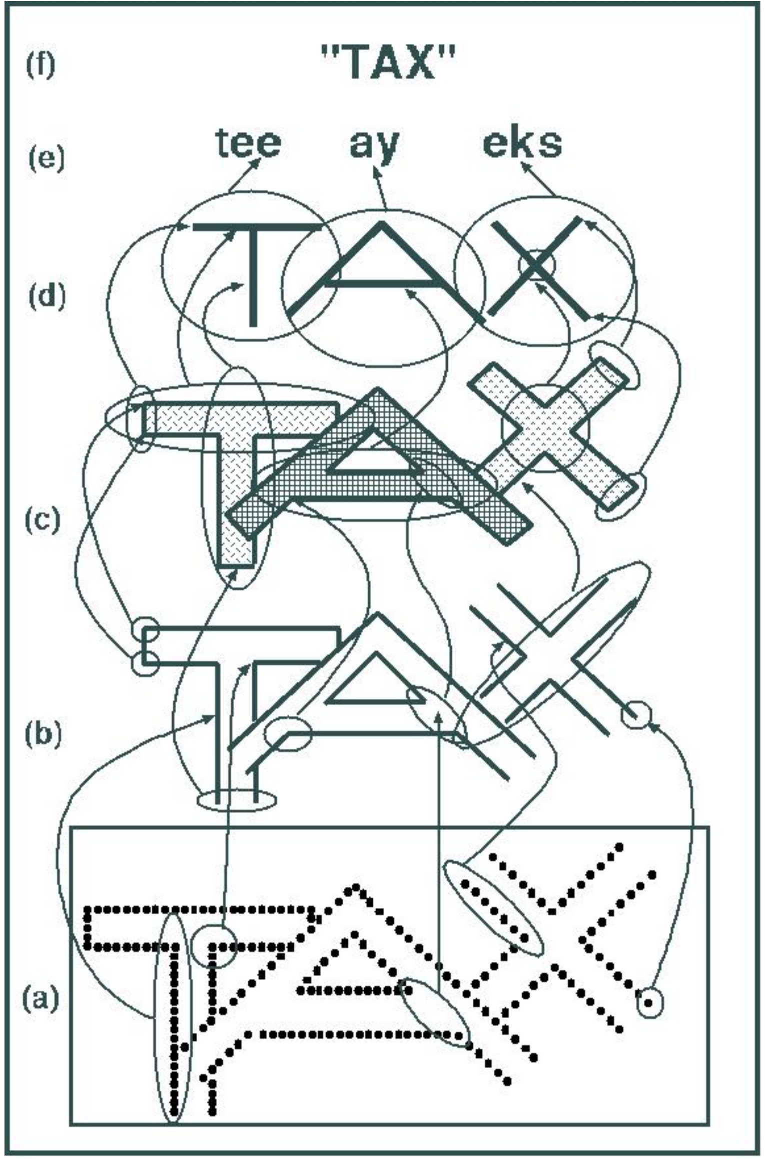
Old conjecture: We process different layers of interpretation in parallel.

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



New conjecture:

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 (and causal) relations within and 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 individual images, rather than claiming to show how to interpret a sequence of images as representing a process.

  Compare R.Grush in BBS 2004

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, ... though it is not 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*)

Ontology available to a visual system can vary

We need to find out more about the ontologies used in vision – and everything else

including both somatic (e.g. sensorimotor – modal and amodal) ontologies
and exo-somatic ontologies.

Ontologies used can vary

- between species
- between individuals in a species

  e.g. being able to read different languages, formalisms, music
- between stages in development of an individual

  learning to see new sorts of things
- between different parts of the same individual

NOTE:

An extension to the ontology need not be definable in terms of what was there before.

New born babies are not born with concepts in terms of which all concepts of modern science can be define.

So we need to explain substantive, not just definitional, ontology-extension.

an old problem in philosophy of science.

(Contrast Fodor: *The Language of Thought*)

Example: Perceiving causation

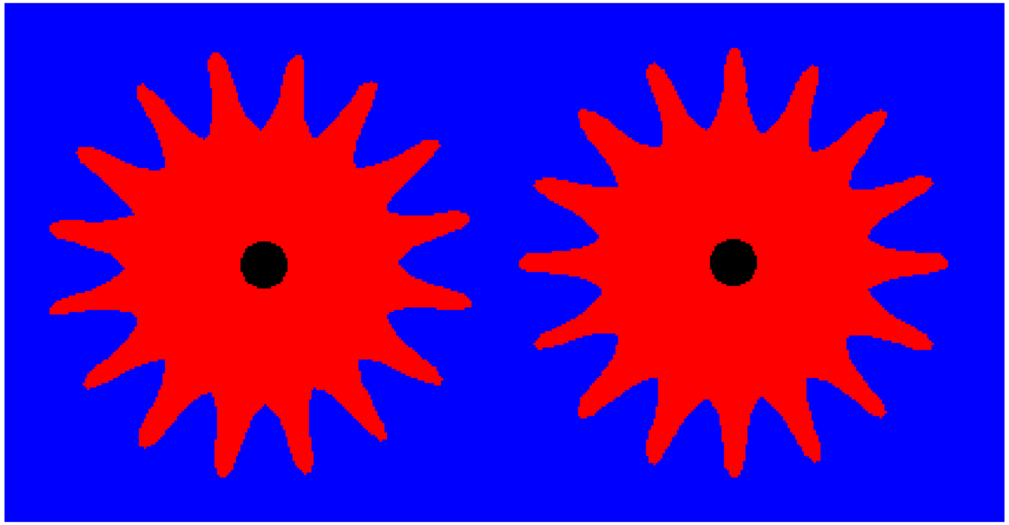
Our ability to perceive moving structures, and our meta-level ability to think about what we perceive, is intimately bound up with perception of causation and affordances.

In some cases the causal relations are inherent in what is seen, whereas in others they involve invisible (hypothesised) structures and processes (as suggested long ago by Immanuel Kant): 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?

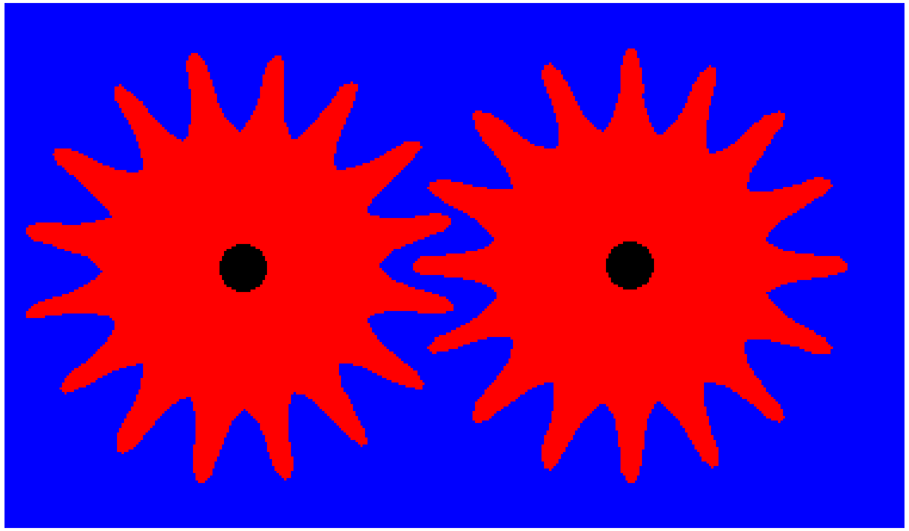
You can tell by experimenting: you may or may not discover a correlation.

Results of experiments to find out which things co-vary under which conditions can be expressed in Bayes Nets (Steven Sloman (2005)).

But you don't really know why the observed causal links exist: they are merely inferred from the statistics (which may include 100% correlations).

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.

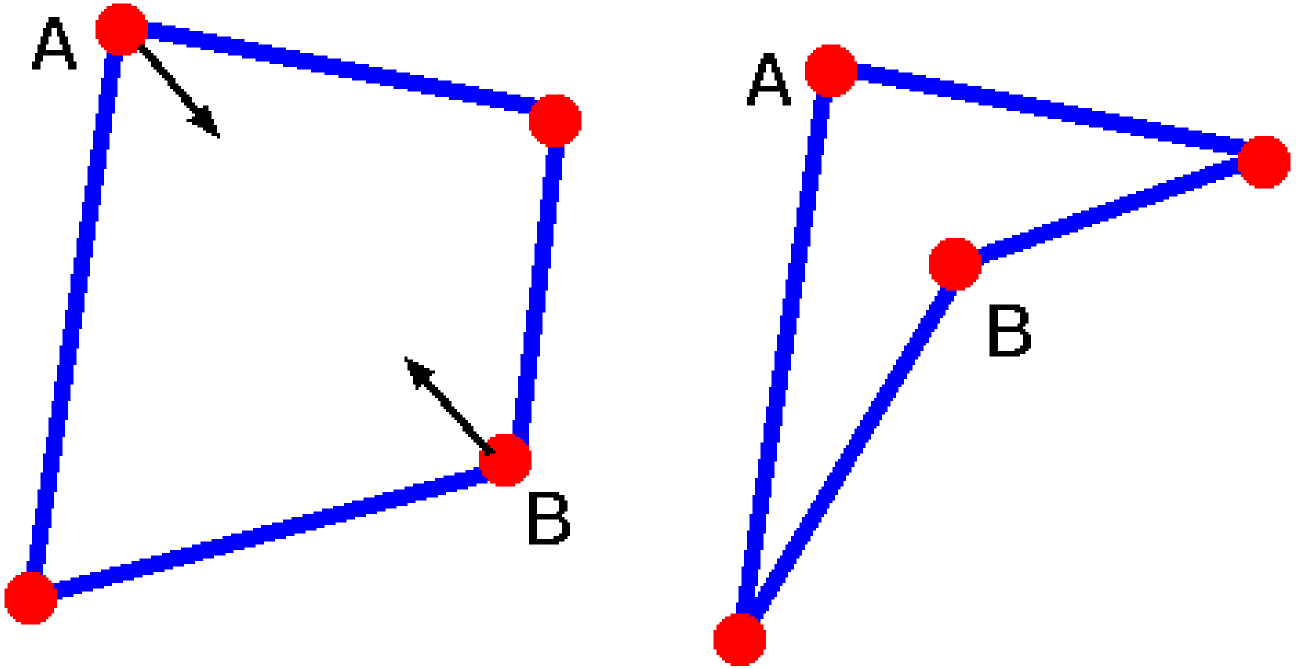
These ideas are developed in two online documents

`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).

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. ???



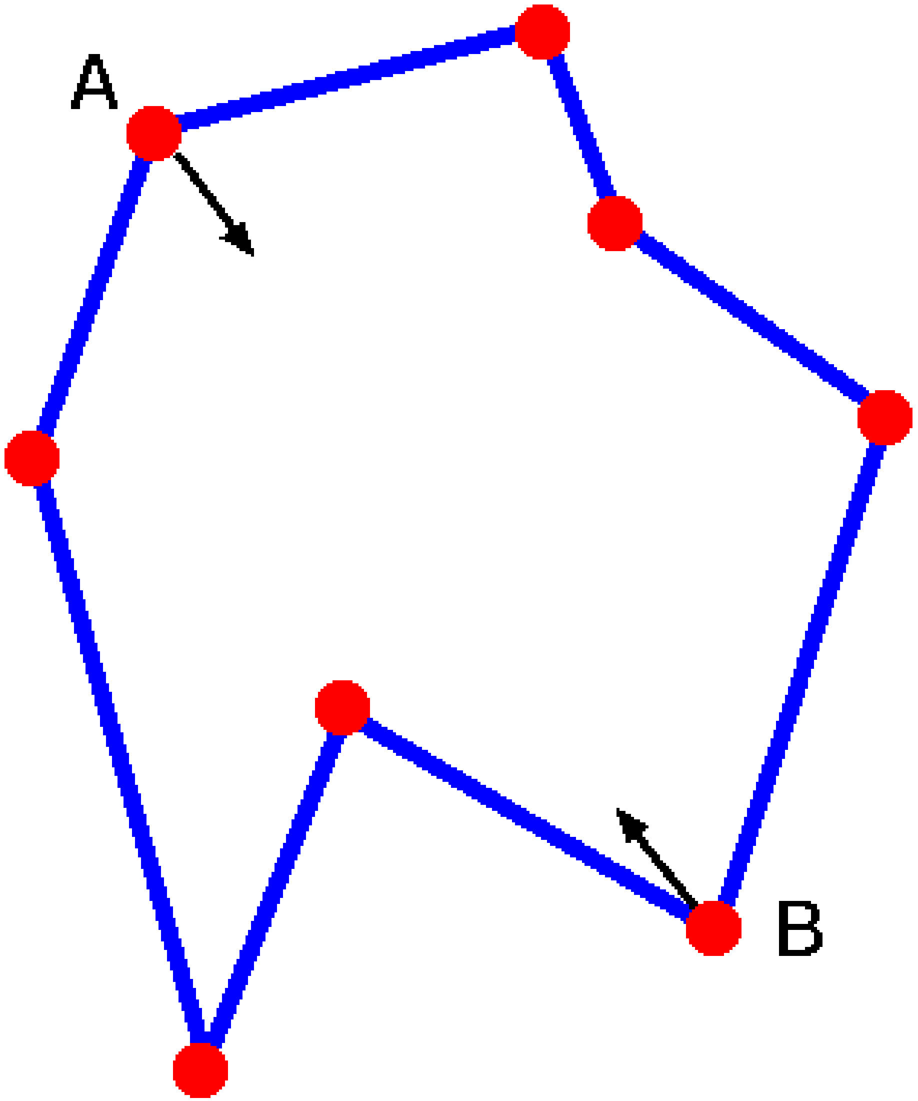
- What happens if one of the moved joints crosses the line joining the other two joints?
- We can change the constraints in our simulations: what can happen if the joints and rods are not constrained to remain in the original plane?

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.



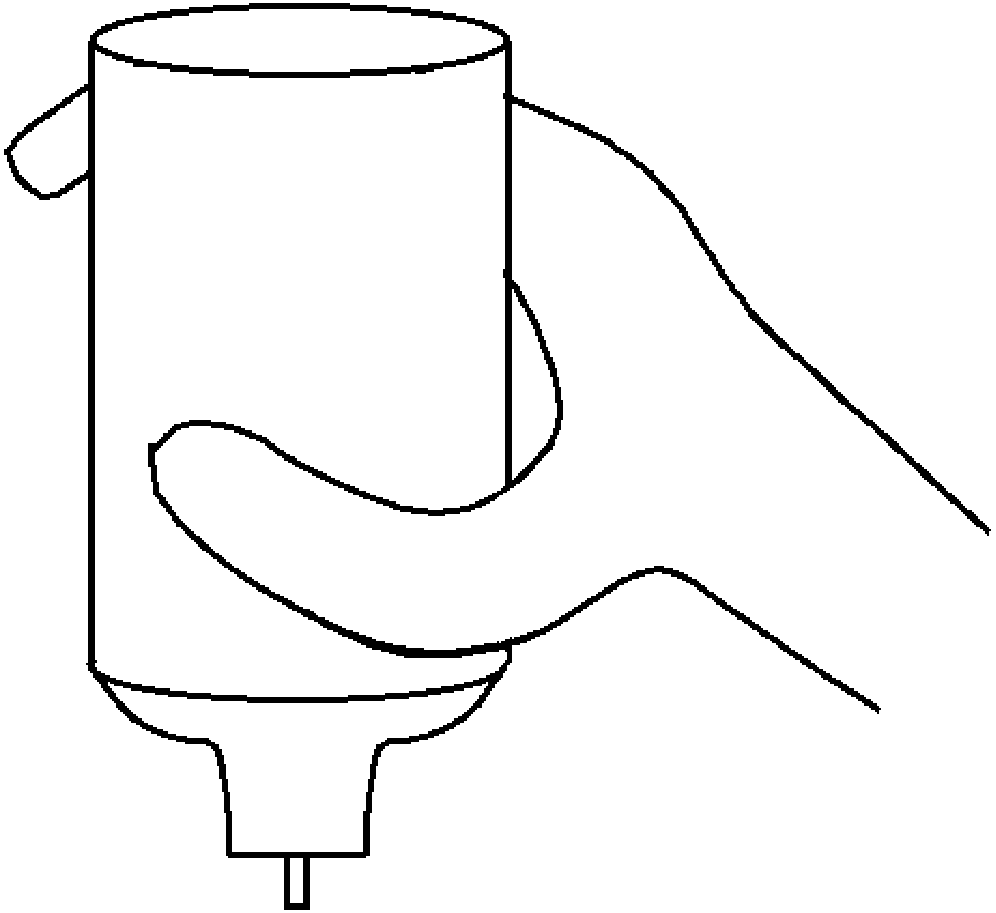
Perhaps intelligent artificial systems will have similar limitations, for similar reasons.

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.

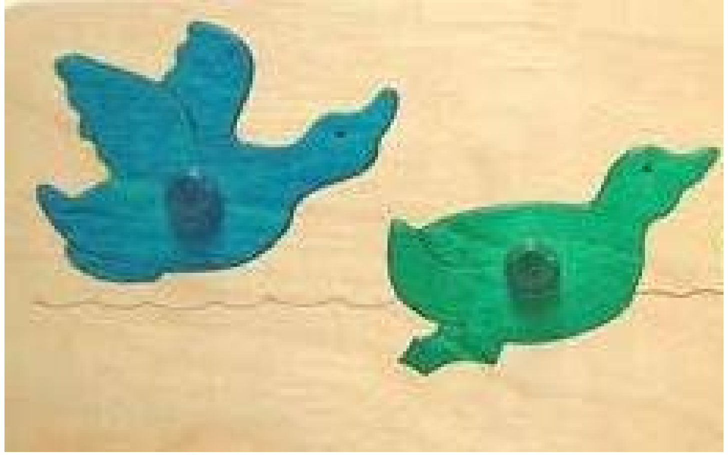
Also, our ability to simulate what is going on explains why as more of the shampoo is used up you have to wait longer before squeezing.

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.

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.

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

The different sub-competences are orthogonal and recombinable:

A feature of what is learnt is that when confronted with a novel configuration (e.g. of gears, levers, strings and pulleys) the representations of each of the sub-processes associated with the parts can be combined in a novel configuration which allows a new simulation to be run to predict or explain how movements in one part of the machine produces movements elsewhere.

This creative recombination is possible because the various parts are in the same space-time region: adjacency in space generates new relationships, unlike conjunction of logical assertions.

A child has to learn to add more and more variety and complexity to what it can simulate, and therefore understand in this way.

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).

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,

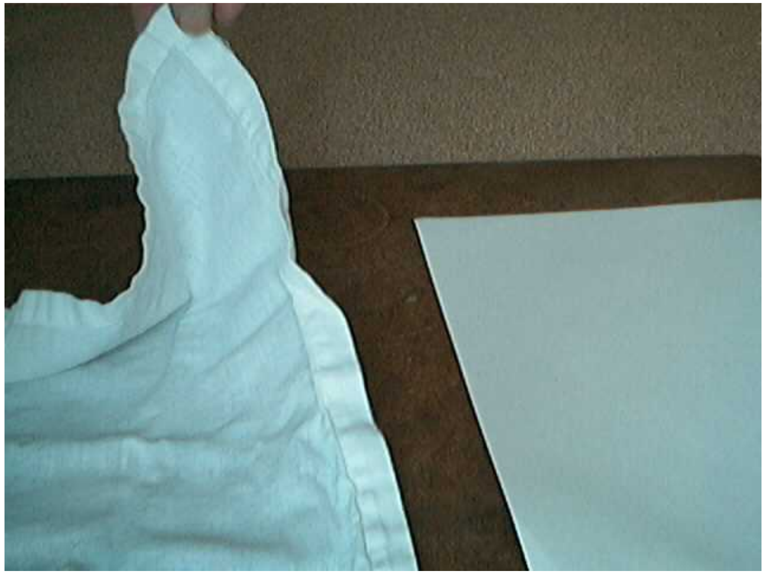
  including, for nest-building birds, twigs, leaves, etc.

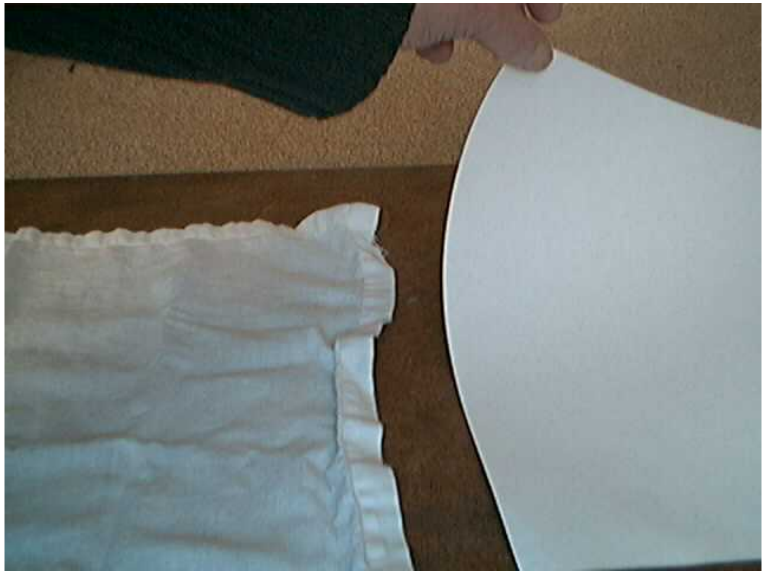
‘Behaviour’ of such things includes their responses to being folded, crushed, picked up, thrown, twisted, chewed, sucked, pressed together, compressed, stretched, dropped, and also the properties of larger wholes containing them.

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.

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.

As a result it is possible to perceive vicarious affordances, i.e. affordances for others, e.g. partly competent infants who may need help, threatening predators or conspecific competitors, etc.

This is essential for many kinds of actions whose goals are to help or hinder other agents.

Constraints on mechanisms
The problem of speed

Some pictures are coming

What do you see and how fast do you see it?

View the pictures at a rate of one or two per second.

Pictures taken by Jonathan Sloman

The problem of speed

The problem of speed

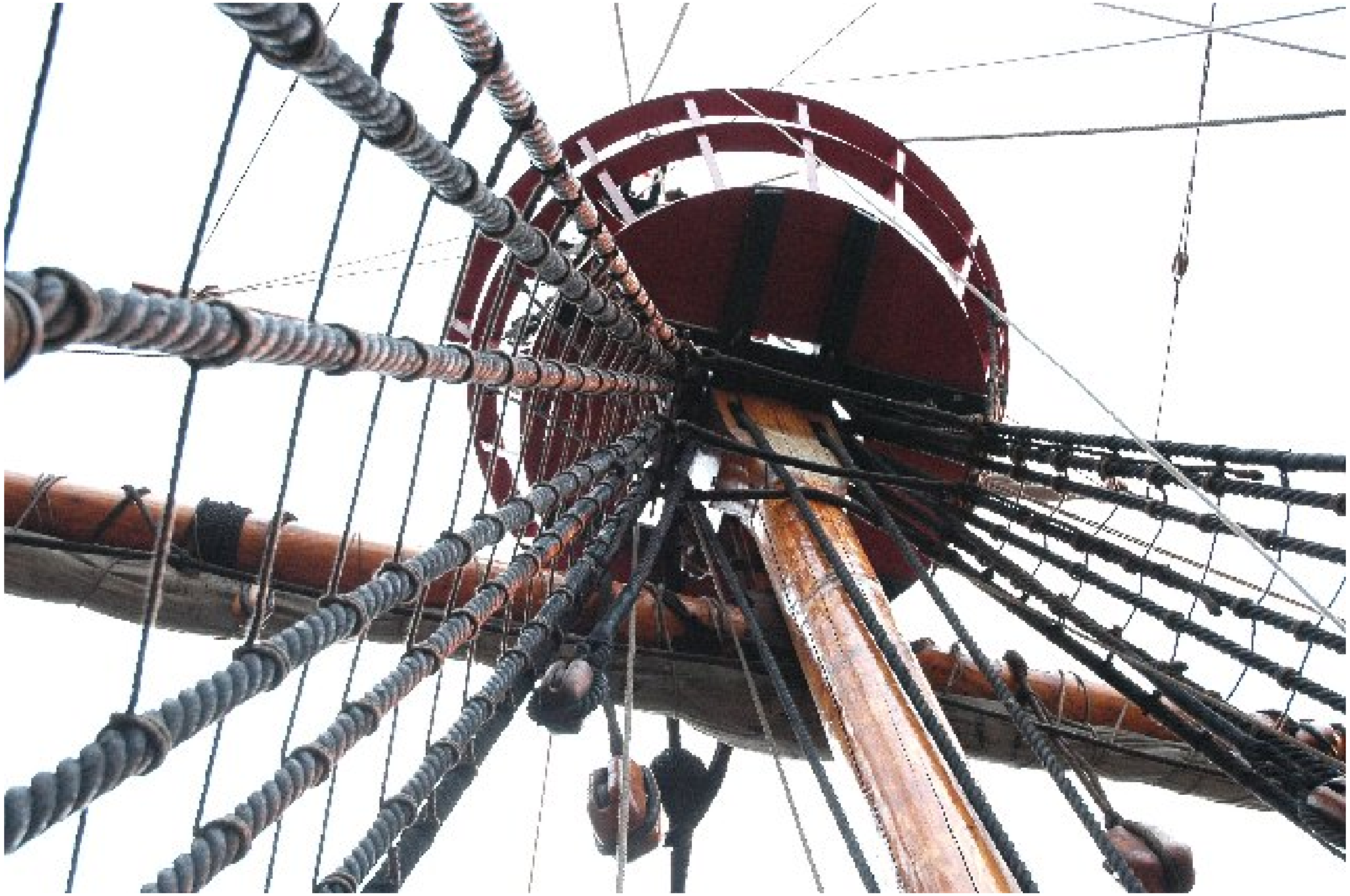
The problem of speed

The problem of speed

The problem of speed

The problem of speed

The problem of speed



In which direction are you looking?

The problem of speed

The problem of speed

The problem of speed

What needs to be explained

The speed with which people can see at least roughly what sort of scene is depicted by each image and what sorts of things are in it implies that our visual mechanisms are capable of finding low level features, using them to cue in features of the image and the scene at various levels of size and abstraction, arriving at percepts involving known types of objects within one or two seconds. Some high level decisions can be made in less than half a second.

This is related to what happens if you go round a corner or come out of an underground station in an unfamiliar town.

I am not claiming that you see everything depicted in all these images, merely that something must be going on to arrive at a 3-D interpretation, including in some cases perceived processes involving pedestrians, bicycles, cars and other vehicles, the state of the weather.

The speed at which this happens, and the variety of types of context in which it can happen, along with the variety of types of information that we can obtain all exceed anything currently possible in computer vision systems.

The inherent ambiguity of all low level image features and the speed with which high level interpretations are formed implies that there is some form of computation going on that defeats combinatorial explosions in the search space of possible interpretations in ways that AI researchers and vision and neuroscience researchers have not yet identified.

A new kind of dynamical system

Perhaps we need a kind of dynamical system

- composed of multiple smaller multi-stable dynamical systems, changing concurrently
- that can be turned on and off as needed,
- some with only discrete attractors, others capable of changing continuously,
- many of them inert or disabled most of the time, but capable of being turned on or off (sometimes very quickly)
- each capable of being influenced by other sub-systems or sensory input or current goals, i.e. turned on, then kicked into new states bottom up or top down,
- constrained in parallel by many other multi-stable sub-systems
- with mechanisms for interpreting configurations of subsystem-states as representing scene structures and affordances, and changing configurations as representing processes
- where the whole system is capable of growing new sub-systems, permanent or temporary, and short-term (for the current environment) or long term (when learning to perceive new things).

This contrasts with

- Dynamical systems with a fixed number of variables that change continuously
- Dynamical systems with one global state (atomic state dynamical systems)
- Dynamical systems that can only be in one attractor at a time
- Dynamical systems with a fixed structure (e.g. a fixed size vector or tree).

Some implications

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.

The information is processed in different ways for different purposes, at the same time using different forms of representation.

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).

Perceptual 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).

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

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.

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, plasticene, 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.

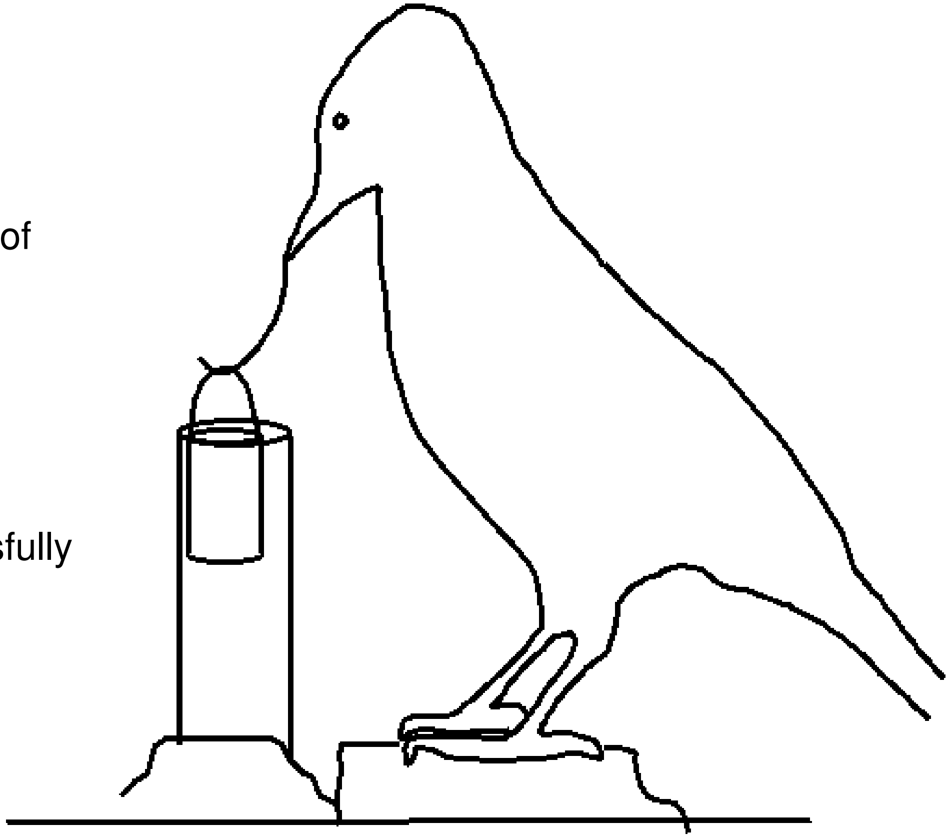
Not only humans

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?

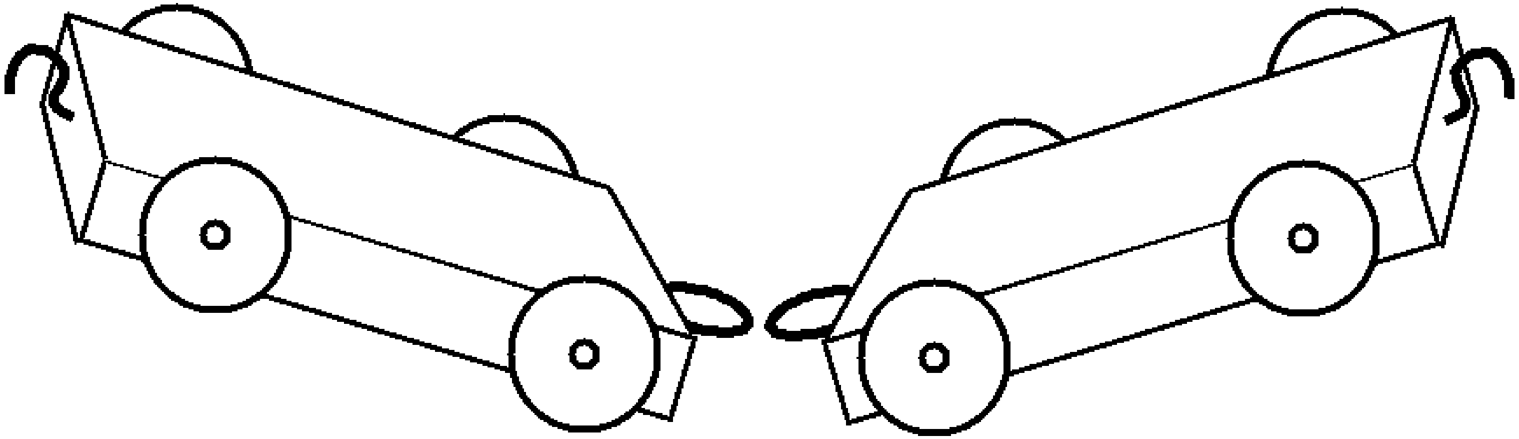
Understanding how hooks work

- 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.

  Understanding how is not a matter of knowing what conditions trigger the change, but what changes within the child to provide the new competences possible. The behavioural data do not determine the answer, since any input-output specification can have infinitely many explanations.
- The next slide points to a video showing a child who has not yet got there.

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

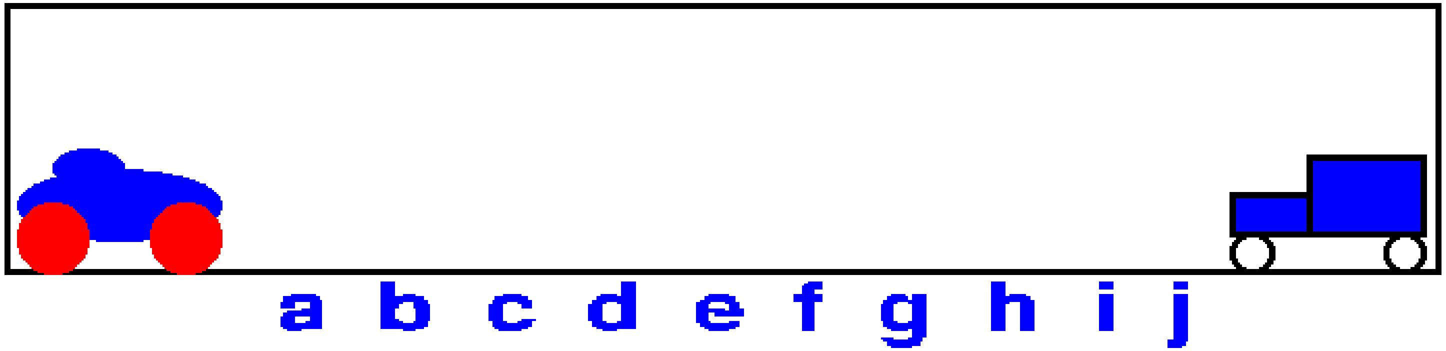
  The date is June 2003, when he was 19 months old. (Born 22 Nov 2001)

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.

Simulating potentially colliding cars



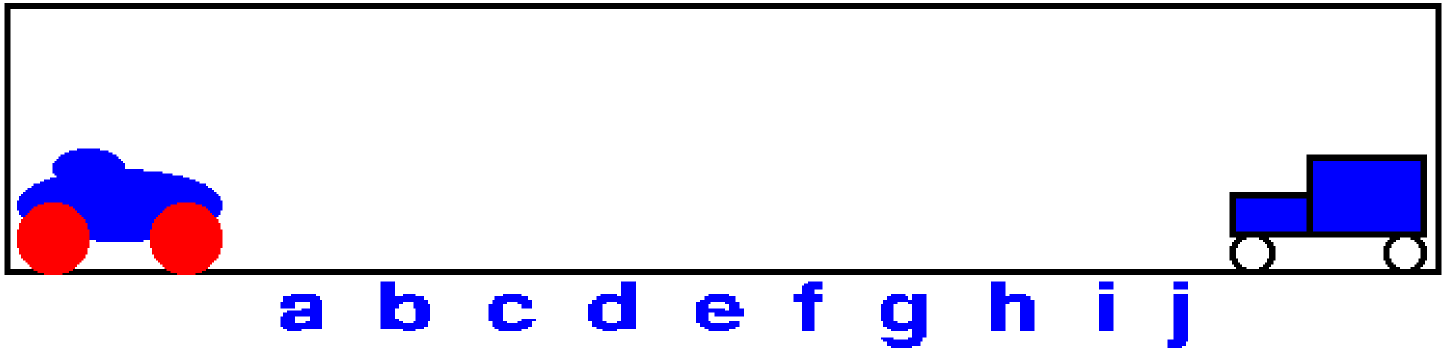
The two vehicles start moving towards one another at the same time.

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

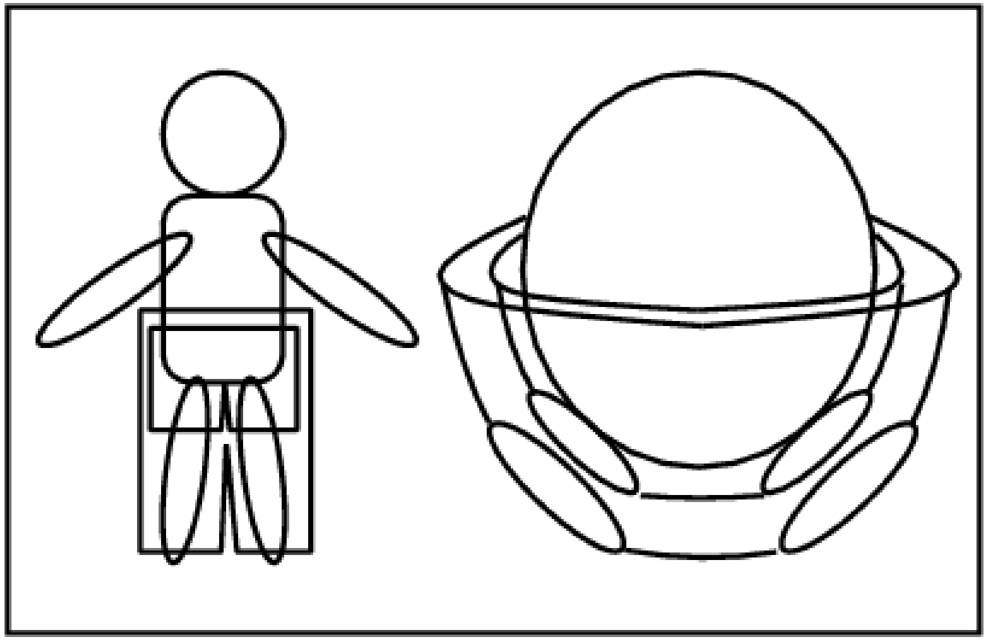
discusses how we can reason about whether Mr Bean (the movie star) can remove his underpants without removing his trousers.
People often don’t see all the possibilities at first.

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).

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, now online) 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:

```
ooo              o         oooo
ooo     +        o    =    oooo
o              ooo         oooo
```

Or is it?
How can it be obvious?
Can you *see* such a general fact?
How?

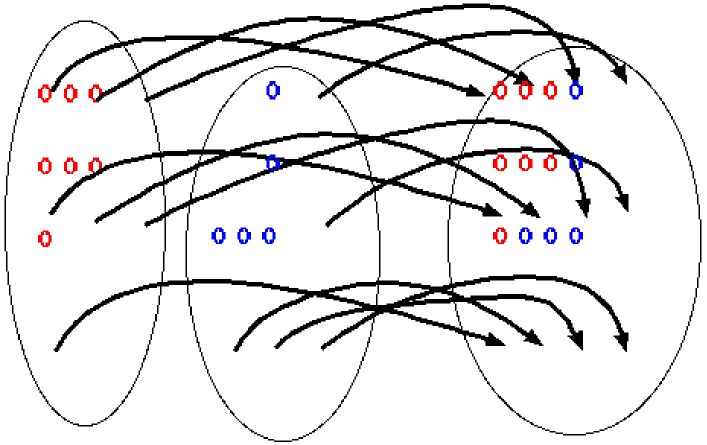
What sort of equivalence are we talking about?

I.e. what does “=” mean here?

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

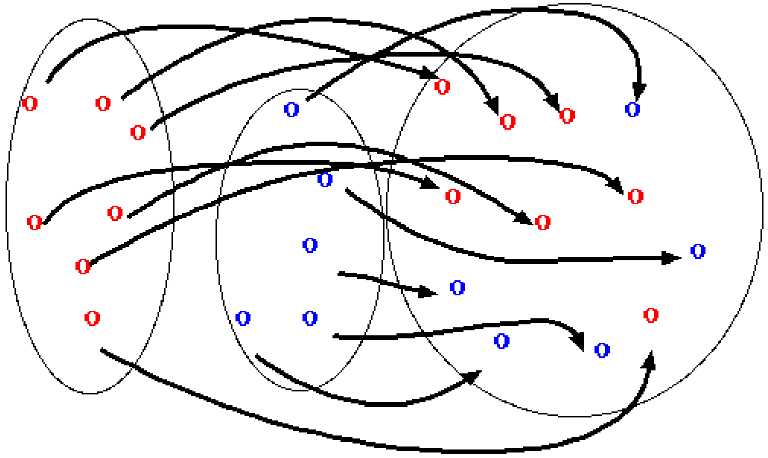
That can be defined logically, but the idea can also be understood by people who do not yet grasp the logical apparatus required to define the notion of a bijection — if they have a way of thinking about the consequences of motion of the blobs.

SEEING that 7 + 5 = 12



Join up corresponding items with imaginary strings.

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?

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

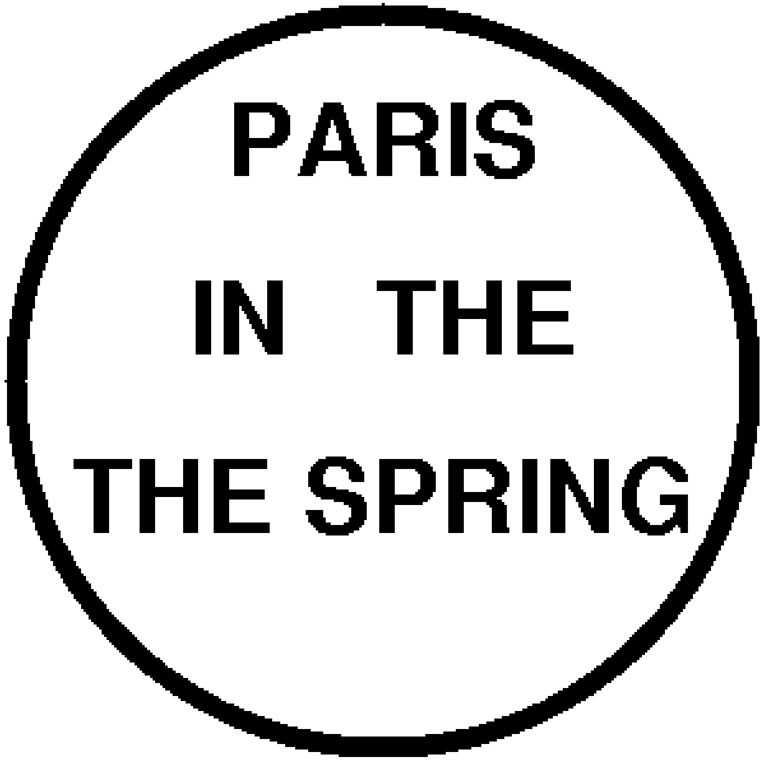
Humans seem to have a ‘meta-level’ capability that enables us to understand why the answer is ‘yes’. This depends on having a model of how our model works – e.g. what changes and does not change if you add another pair of objects joined by a string.

But that’s a topic for another occasion.

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.

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.

The concurrent simulation theory in more detail

What it does and does not say

- 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'.)

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

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.
  Recently made available online `http://www.people.umass.edu/trehub/`

  There are probably lots of things I should have read but have not.

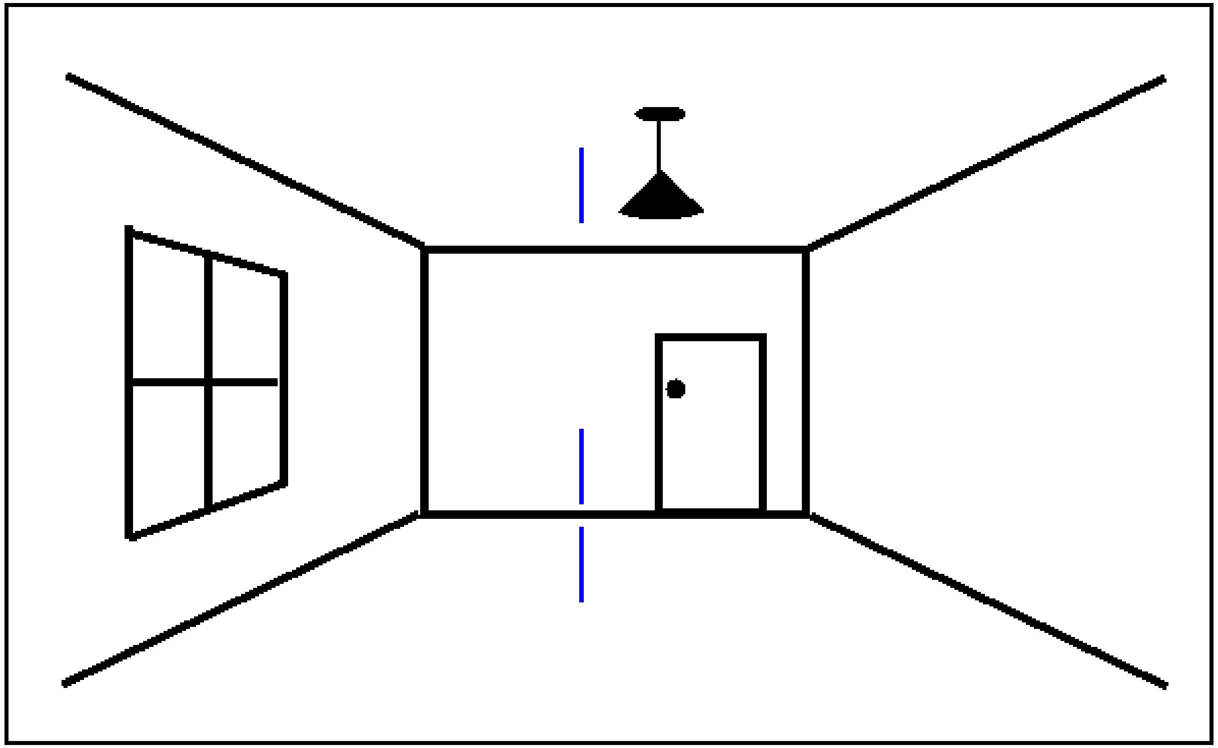
  There is considerable overlap with the BBS paper by R.Grush (2004): The Emulation Theory of Representation.

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.

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 and tracking of instances of object types. However recognition does not require understanding of spatial structure.
- 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/

Some References

Some of my previous work on this topic.

(1971) Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence, in *Proc 2nd IJCAI* Reprinted twice. Online at `http://www.cs.bham.ac.uk/research/cogaff/04.html#200407`

(1978) Chapters 7 and 9 of *The Computer Revolution in Philosophy*, online at `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`

(1982) Image interpretation: The way ahead?, in *Physical and Biological Processing of Images* Eds. O.J. Braddick and A.C. Sleigh. `http://www.cs.bham.ac.uk/research/projects/cogaff/06.html#0604`

(1989), On designing a visual system (Towards a Gibsonian computational model of vision), *Journal of Experimental and Theoretical AI* 1, 4, `http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7`

(2001), Evolvable biologically plausible visual architectures, *Proceedings of British Machine Vision Conference*, Ed. T. Cootes and C. Taylor, BMVA, `http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#76`

(2005a) A (possibly) new theory of vision. (PDF presentation) `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0505`

(2005b) Two views of child as scientist: Humean and Kantian (PDF presentation) `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0506`

(2005c) DR.2.1 Requirements study for representations (Interim report from CoSy Robotic project) `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0507`

(2006) Orthogonal Recombinable Competences Acquired by Altricial Species (Blankets, string, and plywood), Discussion paper, CoSy robotic project. `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0601`

Papers with Jackie Chappell (biologist).

(2005) The Altricial-Precocial Spectrum for Robots (in IJCAI'2005) `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0502`

(2007a) Natural and artificial meta-configured altricial information-processing systems (To appear in IJUC 2007?) `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609`

(2007b) Computational Cognitive Epigenetics (Commentary to appear in BBS 2007?) `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0703`