Originally presented at Birmingham and MIT March 2002
Last revised February 1, 2008

Getting meaning off the ground: Symbol-grounding *vs* Symbol-tethering

(Previously called Symbol-attachment)

Aaron Sloman
School of Computer Science,
University of Birmingham, UK

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

These slides are available in my “talks” directory at:
`http://www.cs.bham.ac.uk/research/cogaff/talks/#grounding`
Shorter newer tutorial version:
`http://www.cs.bham.ac.uk/research/cogaff/talks/#models`
Many talks expanding on these ideas are here:
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`
especially: What is Information?
`http://www.cs.bham.ac.uk/research/cogaff/talks/#inf`
Supervenience, Implementation and Virtual Machines:
`http://www.cs.bham.ac.uk/research/cogaff/talks/#super`

An old, tempting, and mistaken theory

Concept empiricism is an old, very tempting, and mistaken theory, recently re-invented as “symbol-grounding” theory and endorsed by many researchers in AI and cognitive science, even though it was refuted long ago by the philosopher Immanuel Kant (1781).

Roughly, concept empiricism states:

- All simple concepts have to be abstracted from experience of instances
- All non-simple (i.e. complex) concepts can be defined in terms of simple concepts using logical methods of composition.

  E.g. if red and line are simple concepts then red line can be defined in terms of them using conjunction

Symbol grounding theories may add extra requirements, such as that the experience of instances must use sensors that provide information in a structure that is close to the structure of the things sensed.

People are tempted by concept empiricism (whatever it is called) because they cannot imagine any way of coming to understand notions like
red line sweet pain pleasure etc.
except by experiencing instances.

KANT: YOU CAN’T HAVE EXPERIENCES UNLESS YOU ALREADY HAVE CONCEPTS.

Alternatives to concept empiricism and symbol grounding

We'll present some alternatives making use of the following ideas:

- Meanings can be to a considerable extent determined by structural relations between sets of concepts (i.e. theories can determine or at least constrain meaning)
- Sensory links can reduce residual indeterminacy of meaning without being the sole basis of meaning: we call this symbol tethering in contrast with symbol grounding
- Millions of years of evolution can produce individuals that have some concepts from birth

  e.g. precocial species such as deer, that can see and run with the herd shortly after birth
- Genetically determined bootstrapping mechanisms can constrain what is learnt by mechanisms that develop concepts "from experience" –

  i.e. what is learned by interacting with the environment may include some innate and some empirical content, in varying proportions.

  Research is needed on varieties of bootstrapping mechanisms and different kinds and amounts of innate conceptual information that suffice for different sorts of organisms or machines.

Before elaborating on that we need to survey some general ideas about meaning.

Meaning and intelligence

John McCarthy suggested many years ago that if a system can work out a solution to a problem then it must also be able to understand the solution if told it by someone else.

This is not true of everything humans can learn, e.g. learning how to play the violin, learning how to talk, as an infant, etc. But we ignore that for now.

The main point is that understanding is prior to knowing: you can understand a question without knowing the answer - but you cannot know the answer without understanding the question.

> Example: knowing certain facts F1, F2, F3, presupposes having the ability to understand some *expression* of those facts.
>
> That requires understanding the *meanings* used to express the facts.
>
> Knowing the facts does not require the ability to understand a linguistic expression: for the understanding may use purely internal forms of representation, like a cat understanding the difference between a closed and an open door.

Grasping meanings is a precondition for acquiring knowledge.

But grasping meanings is not a simple all-or-nothing matter.
There are many different sorts of meaning-related capabilities.

Common presuppositions regarding meaning

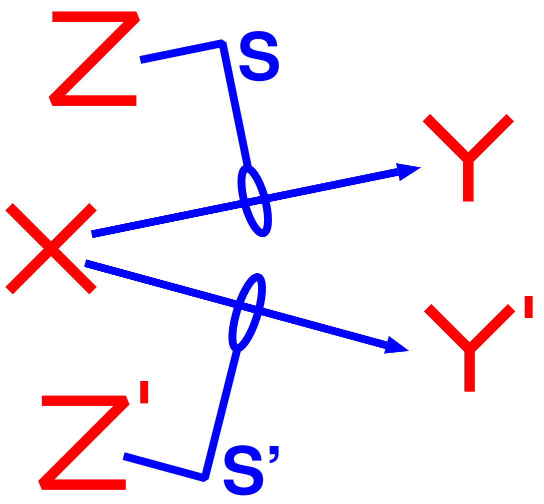
A common model of meaning — three components:

- sign/signifier/symbol/representation
  X
- referent/denotation/signified/extension
  Y, Y',
- user/interpreter/speaker/hearer
  Z, Z',

Some thinkers (e.g. Mill, Frege) added a fourth component:

- sense/connotation/intension: S, S',

Meaning as a three-termed relation
X means Y for Z
X means Y' for Z'



This fourth component is somehow thought to *mediate* the reference to Y, Y'.

This is supposed to be a more direct or immediate kind of meaning, and *must* be grasped for X to be understood: grasp of meaning requires grasp of sense.

However failure of reference is possible, where Y, Y', do not exist, though meaning is understood: e.g. 'The present King of France is bald.'

Other components of the meaning relationship

It is possible to add other components to the meaning relationship, e.g.

- The producer (sender, transmitter, etc) of the information (as opposed to the interpreter)
- The medium or vehicle used used e.g. the language – including
  - the basic components used in its expression e.g. lexical items, pictorial elements,
  - the modes of composition and kind of variability allowed – i.e. the syntax
  - the rules or conventions, implicit or explicit, relating medium to meanings
- The specific context that specifies some aspect of the sense

  (E.g. if someone says 'look for a big stick', the hearer needs to know that what counts as being big depends on the purpose for which the stick is required)
- The social context required for the language to work
- The information-processing mechanisms within language users, required for the language to work.
- The style used (there are many levels of stylistic variation,)
- The purpose for which the meaning is expressed or used
- The emotional content (if present)

These could all be added to the diagrams – though they would become very messy!

Our diagrams and the associated theories are too simple

The diagrams merely express static relationships.

We need a theory of meaning that explains what goes on in a system that does things, e.g. perceives, remembers, refers, wants, evaluates, selects.

I.e. we want a theory that accounts for the role of meaning within a functioning system, such as a biological organism or a robot.

We need something more than the purely *structural* models shown in the diagrams.

A more general framework

We need to talk about "information-using systems" — where "information" has the everyday sense, not the Shannon technical sense. This notion is being used increasingly in biology.

What are information-using systems?

- They acquire, store, manipulate, transform, derive, apply information.
- The information must be expressed or encoded somehow, e.g. in simple or complex structures – possibly in virtual machines.

  (The use of *physical* symbol systems is often too restrictive.)
- These may be within the system or in the environment.
- The information may be more or less explicit, or implicit (e.g. distributed, superimposed).
- The information may be control information

  (e.g. "Do X" rather than "X is the case")

A theory of meaning as we normally understand "meaning" in human communication and thinking should be seen as just one special case within a theory of information-using animals and machines.

Some requirements for a theory of information-manipulators

We need a theory of

- types of information users,
- types of information uses,
- types of information contents,
- types of ways of manipulating information

  For instance:
  - acquiring it
  - storing it
  - comparing
  - deriving
  - transmitting
  - interpreting ....etc....

- ways in which information can control something

Those different topics are all tightly interrelated.

There's more here http://www.cs.bham.ac.uk/research/cogaff/talks/#inf

What are concepts?

People often talk about grasping concepts as somehow being the basis of the ability to understand meanings.

E.g. you understand the meaning of “Lemons are yellow” because you grasp (a) the concepts “lemon” and “yellow”, and (b) the semantic relation expressed by
“A’s are B”.

Concepts can also be used in perceiving lemons, mistaking something for a lemon, wanting a lemon, wondering where lemons grow, looking for a lemon, etc.

Grasping the concept seems to involve something like the ‘S’ in the diagram, since there need not be a referent e.g. understanding:
‘The crocodile in my pocket is hungry’ (There isn’t one.)

The ability to have and use concepts, to understand meanings, to think, want, intend, desire, expect, wonder about — are all aspects of human intelligence involving use and manipulation of information.

Some other animals have some or all of these abilities, though we don’t know which animals have which abilities (or even which ones human infants have).

These are all special cases of the more general class of abilities shared by all biological organisms: the ability to use information.

Talk about having or understanding concepts is a short-hand for reference to an immensely complex collection of capabilities.

“Structural” models (e.g. meaning is a 3-termed or 4-termed relation) say nothing about mechanism

- There are diagrams and theories indicating how many things are involved in meaning/signifying/expressing/referring.
- But these are static models, lacking in explanatory power.

  They say nothing about how a grasp of meaning plays a causal role in processes such as
  - perceiving a situation,
  - learning a generalisation,
  - becoming puzzled about something,
  - finding a solution,
  - applying previously learnt information,
  - desiring something,
  - working out how to get what one desires,
  - storing information and later finding it again,

  ... etc.
- A theory showing how a grasp of meaning is put to work, will have to be a theory of mechanisms interacting in some architecture.

The simple and intuitive concepts crumble into confusion if examined closely.

It proves quite difficult to define these ideas about meaning and concepts precisely so as to cover all cases in a clear way.

- Does a proper name (e.g. "London") have a sense (S) as well as a reference (Y)? Mill said "No". Frege said "Yes". Then found that he had trouble with that idea. What is the sense of "London"? Or the sense of "I" or "you" or "we"?
- When I see a car coming towards me as I start to cross the road, and I therefore step back off the road, am I using concepts
  - of a car,
  - of motion towards me,
  - of continuing to walk, or moving out of the way.

Does it depend which part of my mind is in control:
the old reactive subsystem
or a newer deliberative layer?

See other talks at http://www.cs.bham.ac.uk/˜axs/misc/talks/

Perhaps all the cases of concept use studied by philosophers are special: they include only cases of self-conscious concept use, monitored by a meta-management/reflective sub-mechanism?

What about other uses of concepts – in reactive sub-mechanisms, or in insects.... ?

What about a fly?

How wide-spread is grasp and use of concepts?

- When a fly sees the fly-swatter approaching and escapes, it acquires and uses information: is it using concepts?
- What about a human infant reacting to a finger touching its cheek by turning towards it and starting to suck?
- What about new-born deer or calf that walks to its mother's nipple and starts to suck? Does it have a concept of the intervening space that it has to traverse?

Perhaps debates about whether machines can understand concepts are a manifestations of a general lack of clarity of the idea of varieties of information and how they can be processed.

Some philosophers discuss a distinction between "conceptual content" and "non-conceptual content".
Normally philosophers do not discuss mechanisms that make the use of meaning or content possible, and therefore do not explore an adequate variety of cases.

Hence the attempt to define yet another dichotomy.

We can regard that as just a naive first step towards a more general theory of the sort we need to find, which will allow more than two types of contents.

Philosophical debates about meaning and AI

AI theorists and philosophers often debate whether computer-based machines can really *understand* anything, i.e. whether they can *grasp* meanings themselves, as opposed to merely appearing to.

Can machines play the role of Z, the symbol user?

- Many AI researchers say: Yes — look at these working systems that understand the symbols they use.
- McDermott warns us: Just because a subroutine is named "understand", that does not make it an understander (Artificial Intelligence meets natural stupidity (1976))
- Searle says: NO – computers, like books and calculators. have only "derivative intentionality" not "original intentionality". (Haugeland's terminology.)
- Penrose says: ONLY if they use super-Turing computers based on as yet undiscovered quantum-gravity mechanisms.
- Harnad says: ONLY if they satisfy conditions for "symbol grounding".
- Dennett says: we are free to adopt the "intentional stance" toward machines (i.e. treat them as having beliefs, desires, intentions, percepts, etc.) if we find that useful, e.g. in predicting their behaviour.

How can we choose between these answers? Maybe we shouldn't!

A different approach

Different architectures support different sorts of meanings grasped in different ways.

Let's try to understand the variety of architectures and the variety of types of understanding
made possible in different sorts of architectures
instead of trying to answer
ill-posed Yes/No questions.

We can attempt a comparison of many kinds of information-users attempting to see how many varieties there are, and how they are similar, and how they differ.

In particular, let's NOT restrict ourselves to adult humans and their introspections.

An information-processing architecture includes:

– forms of representation,
– algorithms,
– concurrently processing sub-systems,
– connections between them

It need not be a rigidly fixed system: some architectures can modify themselves, e.g. a unix system that can spawn new processes that can spawn new processes, or a child's mind.

We need to understand the space of information-processing architectures ("design space") and the states and processes they can support, including:

– The variety of types of perception
– The variety of types of reasoning
– The variety of types of emotions
– The varieties of types consciousness
– ...

For more on information and information-processing virtual machines see

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

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

What an organism or machine can do with information depends on its architecture

Not just its physical architecture – its information-processing architecture.

This is typically a virtual machine with components that are virtual machines, operating on structures in virtual machines, like lists, graphs, numerals, distributed processing networks, not neurons, chemicals, electrical signals.

Familiar virtual machines include

- a chess virtual machine
- a Lisp or Prolog or Java virtual machine
- a word processor
- a spreadsheet
- an operating system (linux, solaris, windows)
- a compiler
- a neural net (implemented on a computer)
- most of the internet

Explaining in more detail what virtual machines are, how they relate to physical machines, and how virtual machine events can have causal powers, is a different talk.

(http://www.cs.bham.ac.uk/~axs/misc/talks/#super)

So we need to ask how various architectures support various meaning-related capabilities

Before doing that let’s look at some philosophical theories about where meanings come from, or how meanings are grasped, starting with concept empiricism.

Concept empiricism

Many people believe that the concept of understanding something is very clear, and we all know what we are talking about when we say Z understands X.

Such people, whether they study philosophy or not, are often intuitively attracted to the ‘concept empiricist’ philosophical theory, i.e. the theory that:

> It is impossible to understand some concept unless
>
> EITHER
>
> you have abstracted the concept from experienced instances of it,
>
> OR
>
> you grasp an explicit definition of it in terms of concepts that satisfy this (recursive) requirement.

E.g. if you think it is impossible for someone who has been blind from birth to understand ‘red’, ‘green’, etc. then you are probably a concept empiricist.
You probably think you learnt colour concepts by experiencing colours.

Distinguish

- concept empiricism (concepts come from experience).
- knowledge (or judgement) empiricism (knowledge comes from experience).

(Neither implies the other.)

Why is concept empiricism so appealing?

Because there appear to be only two ways of learning new concepts:

- EITHER You are given a definition of a new concept in terms of old ones
  (E.g. 'prime number' defined in terms of division and remainder)
  (Requires some formal apparatus for expressing definitions.)
- OR You are shown examples, and then you somehow get the idea.

  Two sub-cases of the second case:

  – You can articulate the definition you have extracted from the examples
    (E.g. Winston's program learning about arches, houses, etc. — amounts to guessing a definition of the previous type.)
  – What you learn is not definable in terms of prior concepts using any formalism you can articulate
    (E.g. learning to recognise Blair's face by seeing it, or learning the taste of kiwi fruit from examples. Chunking in self-organising neural nets.)

In both cases we have no introspective insight into the processes that produce the new concepts. People merely assume that we can somehow (by magic?) store something that is abstracted from the individual experiences.

There are many implausible theories of meaning produced by philosophers and psychologists who don't know how to design working systems: e.g. "We remember all the instances and use a similarity measure"; "We store prototypes and compare them with new instances", etc., etc.

Symbol grounding: a form of concept empiricism

Harnad wrote in his 1990 paper:

"...there is really only one viable route from sense to symbols: from the ground up..."

Our alternative suggestion:

- Many kinds of meaning are constructed in an abstract form, determined by the structures and processes that can occur in the system using them.
- This always leaves meanings partly indeterminate.
- Meanings can be made more definite by associating the theory with transducers and specific modes of interaction with instances of the concepts (where that makes sense):
  I.e. existing, un-grounded meanings are made more definite by becoming "tethered"

I.e. it is better to think of such meanings as constructed "top-down" via a process of successive refinement and symbol tethering.

The whole apparatus may have been constructed bottom-up by evolution, but that is irrelevant: how the machinery came to exist is not relevant in a description of what it *does now*.

A similar system produced all in one go by a designer or some random sequence of assembly events would support the same states and processes, despite having no evolutionary origins. (R.A. Young, Proc. Aristotelian Society, 1994)

Kant's refutation of concept empiricism

The philosopher Immanuel Kant attacked concept empiricism on the grounds that you cannot have any experiences at all without already having concepts used in the experience, so those concepts cannot have come from experience.

E.g. the notion that you can use experiences of triangles to abstract the concept "triangle" presupposes that you can have spatial experiences — usually that presupposition is never fully analysed. (We'll start analysing it below.)

Other problems for concept empiricism

- Concepts of unobserved entities
- Theoretical concepts in scientific theories
- Concepts of dispositions that are not realised
- Logical and mathematical concepts (e.g. "not")

All discussed by philosophers, e.g. philosophers of science.

Let's focus on spatial concepts, in Kantian fashion.

What are spatial experiences?

- Spatial experiences require understanding of a spatial *region* in which there are spatial relationships (above, below, enclosed, connected, etc.)

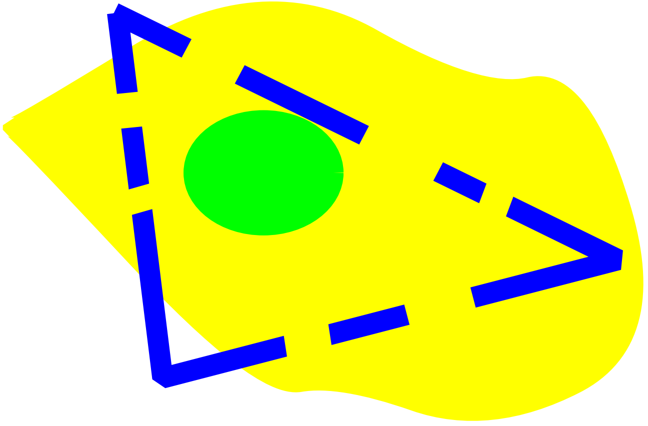
  Do different animals experience different spatial relations?

- Likewise experiencing colours requires having spatial percepts involving extended surfaces capable of having properties other than the geometric ones. (E.g. colour, texture, hardness, smoothness, warmth, etc.)

What is involved in experiencing a triangle?

- segmentation including figure/ground separation.
- grouping fragments
- detecting collinearity, straightness, etc.
- detecting global features: closure, number of sides, etc.

etc.



What must happen inside a system capable of experiencing triangles?

Do we have any idea?

We do not experience what it is to have an experience!

What is experiencing an empty space?

Consider Picasso looking at such a space.

Seeing an empty space is being *potentially aware* of a large variety of structures and processes that could occupy some of that space.

An example of the important notion of seeing affordances.

What kind of information-processing architecture makes that possible? How might it display its grasp of empty space?

Minsky's question: could a piece of string have only one end?

His young son's answer: It would have to fill the whole universe.

What sorts of mechanisms supported his thinking about the question?

Aspects of the perception of physical objects

Seeing physical objects as we do includes

- Segmenting things in 3-D space
- Understanding that objects have surfaces
- Understanding that surfaces can have extended properties that vary over the surface, e.g.
  - colour
  - warmth
  - texture

  etc.

What sort of information-processing architecture is required to support that way of conceiving of objects?
Is there a unique answer? Do all animals do it in the same way?
Do all parts of the human information-processing architecture do it in the same way?

Is it done in the same way for visual perception and for tactile perception?

Is it done in the same way by deliberative and reactive mechanisms?

Is it partly the same and partly different? how?

More on spatial experience

Many of the requirements for visual spatial experience overlap the requirements for tactile/haptic spatial experience.

This follows from the fact that the structures experienced visually and the structures experienced through touch and movement have much in common.

A good engineer designing a seer and a feeler would ensure that visual and tactile spatial perception produced an *integrated* information structure.

Would you trust a dentist without integrated visual and tactile spatial perception?

Neo-Kantian Conjecture:

- Evolution produced brains (for humans, chimps, ...) which have a powerful grasp of space and motion (i.e. there are information-structures and mechanisms for operating on them, which can store and manipulate information about spatial objects and their properties
  – probably in special-purpose virtual machines.)
- The mechanisms are partly underdeveloped at birth and bootstrap themselves through action.

  (For precocial species, e.g. sheep, deer, horses, chicks, etc. they are closer to their final state at birth than for altricial species: humans, lions, monkeys, eagles.... Why?)
- Within this framework spatial objects have surfaces with *extended properties* of various sorts: texture, hardness, curvature, stickiness, colour, temperature, ...
- The conceptual apparatus for colour and for texture share a huge amount.
- People blind from birth have and use that apparatus.
  They have most of what's required to understand talk of colours.

A multitude of spaces

Do a midge, a mouse, a monkey and a man see the same space?

We do not assume that space is something uniform and understood in the same way by all animals with spatial capabilities: there are important differences that are ignored in the mathematical characterisation of Euclidean space.

We need to allow that an organism's understanding of space will be related to

- what its spatial sensors can do
- the kinds of goals it can have
- the kinds of actions available to it
- the information processing capabilities that relate sensor information to possible goals and possible actions: i.e. in perceiving affordances.

So spatial concepts used by an individual Z will related to kinds of motion Z can perform, the kinds of objects Z can manipulate and how it can manipulate them, the kinds of spatial planning and plan-execution Z needs to be able to do.

Different spatial concepts will be related to different domains of activity for the same individual Z, e.g.

- visible manipulable 3-D objects
- regions (like a room) in which almost everything is visible from every where,
- regions (like a house) in which most things are not visible from most places
- larger scale regions, like a town or a country where relevant actions (e.g. travelling) may be extended over long periods of time.

All such concepts are probably partly a result of evolved innate constraints.

Meaning without experience

If all the above is correct then a congenitally blind person with a rich grasp of the nature of space and extended surfaces already has most of what it takes to understand colour concepts.

Exactly what such individuals have still needs to be analysed: that requires us to develop a detailed theory of the architecture, the forms of representation, the information-processing mechanisms they have available – a long term multi-disciplinary investigation.

The mechanisation of understanding

The ability to understand is rooted NOT MERELY in a history of personal experience of instances but in a collection of mechanisms for manipulating information of the appropriate structure.

- Different mechanisms in different architectures will support different varieties of understanding, and different varieties of uses of meanings.
- The mechanisms will create and manipulate structures (usually not ‘physical symbols’ but symbols in virtual machines).
- But the ability of structures to encode meanings, to store or convey information, will depend in part on the intrinsic properties of the structure and in part on the mechanisms creating, manipulating and using the structures.
- Examples are pictures, collections of sentences, collections of logical formulate, and many others.
- What makes it possible for such things to be used as information structures?

A key idea

A formal system determines a class of possible interpretations (models) and thereby gives the undefined symbols within it a (somewhat indeterminate) meaning.

Example: axioms for lines and points (projective geometry)

Any two lines will intersect in exactly one point.

Any two points will be joined by exactly one line.

P1, P2, P3 are points.

L1, l2, L3 are lines.

L1 joins P1 and P2. L2 joins P2 and P3. L3 joins P3 and P1.

P1 is on the intersection of L3 and L1
P2 is on the intersection of L1 and L2
P3 is on the intersection of L2 and L3

Models/interpretations for meaningful structures

There can be many complex structures whose parts and their relations form a model or interpretation for the above collection of statements.

Ambiguity is possible:

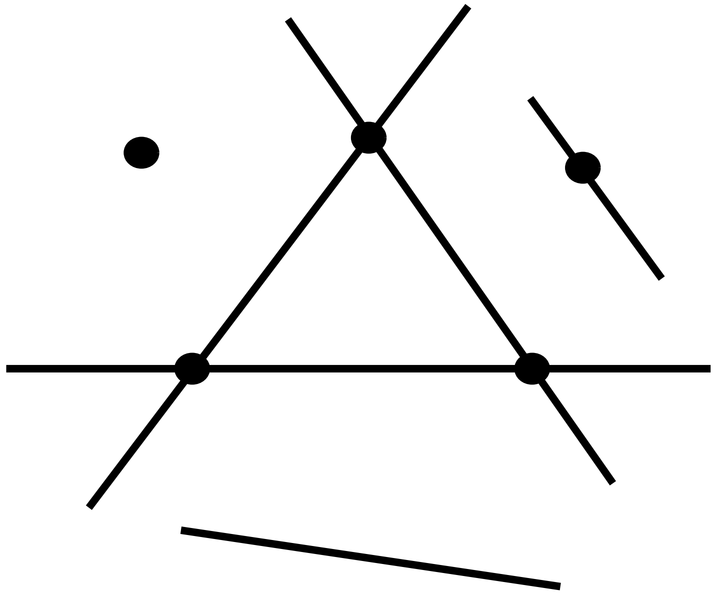
- It is well known that a picture can be ambiguous (e.g. necker cube, duck-rabbit).

- However a logical specification such as the above can also have alternative ways of being understood.

Dual interpretations

In any space with lines and points containing a model of our set of axioms there will be a dual model, with lines and points switched.

Consider a collection of lines and points (not all shown in the figure) satisfying the axioms. It is a well known fact of projective geometry that the axioms of our earlier slide can be mapped onto such a collection in two different ways.



The *non-obvious* way is to treat the points as referred to by the symbol “line” and lines as referred to by “point”.
Then the relations between lines and points have to be swapped.
(E.g. “any two lines intersect in one point” becomes “any two points are joined by one line”.)

This illustrates an important general point: a complex meaningful structure may be capable of being mapped onto some complex part of the world in more than one way

However there is not usually such a dramatic duality!

In this example, the abstract structure and all the operations we can perform on it define a kind of "generic" meaning

The generic meaning, which limits the class of possible models, can be made more determinate in at least two different ways:

- by attaching (how?) some of the symbols to individuals or classes of individuals, or features, in some portion of the world
- by adding new non-redundant "axioms" (content expressions)

Both methods presuppose that sophisticated information-processing capabilities exist in whatever system is doing the work.

Linking parts the whole system to the world reduces the variety of interpretations and thereby increases the determinateness of the meanings of the primitive symbols.

But interesting conceptual systems always remain partly indeterminate and capable of further development – this is part of what makes the continual growth of science possible: old concepts enter into new relationships, sometimes with new primitives (via 'meaning postulates').

Reducing semantic indeterminacy

An architecture that builds and manipulates structures may interpret undefined "primitive" symbols (predicates, relations, functions, etc.) as somehow referring to components of reality.

However, as indicated in previous sections the initial construction may not suffice to specify with complete precision what is referred to.

The precision can be increased, indeterminacy decreased, in various ways. Adding new tethering points increases the semantic determinateness of the "undefined" symbols. (Similar ideas can be found in the work of Quine: e.g. in his 'Two dogmas of empiricism').

We need much research to investigate the many ways in which this can work, within different sorts of architectures with different sorts of mechanisms, e.g. logical mechanisms, self-organising neural mechanisms of various sorts, etc.

The strengths and weaknesses of a wide range of mechanisms and architectures need to be understood. We shall then be in a far better position to propose good theories of concept formation in children.

In particular, it will be interesting to see what difference it makes if the architecture includes a meta-management (reflective, self-referring) component. (As in recent work of Minsky and myself.)

Understanding colour concepts

From the previous conjectures it follows that grasping colour concepts may simply be a matter of plugging certain parameters (perhaps procedural parameters) into a complex mechanism — where alternative parameters are required for other concepts of properties of extended surfaces, e.g. texture, warmth, etc.)

- So someone blind from birth may be able to *guess* (not necessarily consciously) parameters to be supplied to produce a new family of concepts of properties of extended surfaces.
- It may even be the case that evolution provides some of those parameters ready made even in blind people who are not able to connect them to sensory input.
- If so, congenitally blind people may be able to share with sighted people most of their understanding of colour concepts: perhaps that is why blind people are often able to talk freely about colours.

POSSIBLE OBJECTION: their understanding will be only partial

REPLY:

The fact that innate mechanisms suffice for such partial understanding of a family of concepts shows that not all concepts need to be derived from experience of instances.

And when experience plays a role it may be a small role!

Kant's refutation – continued

Kant argued, as indicated above, that there must be innate, *a priori* concepts,

- not derived from experience,
- though possibly somehow "awakened by" experience.

He thought the operation of concepts was a deep mystery, and gave no explanation of where apriori concepts might come from, apart from arguing that they are necessary for the existence of minds as we know them.

The now obvious option, that apriori concepts are products of evolution, was not available to Kant.

HAD HE LIVED NOW, HE WOULD HAVE BEEN DOING AI.

Other philosophers tend to treat "having an experience" as a sort of unanalysable, self-explanatory notion. From our point of view, and Kant's, it is a very complex state, more like a collection of processes, along with a large collection of dispositionally available additional processes, all presupposing a rich information-processing architecture, about which we as yet know very little.

Conjecture

Evolution produced “meaning-manipulators”, i.e. information-processing systems (organisms) with very abstract and general mechanisms for constructing and manipulating meanings.

These mechanisms are deployed in the development of various specific types of meaning, some provided innately (e.g. for perceptual and other skills required at birth) and some constructed by interacting with the environment. (Some species have only the former.)

These mechanisms provide support for the organism’s *ontology*: the collection of types of things supposed to exist in its environment and the types of processes that can occur, etc., including actions.

Specific contents for some of the generic ontologies can be provided by linking to sensors and motors. But even without that there is a generic grasp of the meaning insofar as the *structure* of a family of related concepts is grasped.

So meanings are primarily created by the mechanisms that manipulate them: the links to reality added through transducers merely help to refine partially indeterminate meanings.

(This is a refinement of the theory in my papers in IJCAI-85 “What enables a machine to understand?” and ECAI-86 “Reference without causal links”.)

McCarthy on the Well Designed Child

McCarthy has a paper addressing many of these issues, though from a slightly different standpoint:

- he is concerned with how to produce a well-designed baby robot rather than trying to understand what biological evolution produced
- however he does make connections between the two, including attempting to identify some evolved but possibly sub-optimal aspects of human infants
- he explicitly asks what *knowledge* should be innate, and only implicitly addresses the question what *concepts* (or conceptual apparatus) should be innate
- he does not discuss architectures and requirements of different components.

John, McCarthy, (1996). The well-designed child. (Last updated 1999)
http://www-formal.stanford.edu/jmc/child1.html

That paper includes a sample list of some of the kinds of knowledge a well designed child might be programmed to use or to look for, instead of being left to do everything bottom up, driven by data, as empiricists would claim.

Empiricist philosophy and empiricist AI

Both empiricist and rationalist (nativist) philosophers have typically never tried to design a working information-processing system.

- Empiricist philosophers just assume that totally general learning mechanisms can be driven by data (by experience) to learn everything that's interesting or useful about the environment.
- Empiricist AI (unwittingly) reinvents empiricist philosophy, and attempts to build such mechanisms.
- That has not achieved much, except in very limited contexts, e.g. learning to cluster data in some fixed dimensional space, or using data to induce correlations between induced categories, or using genetic algorithms or genetic programming to evolve interesting solutions to specific problems.

Rationalist AI

Crude rationalist AI starts from the assumption that some single very powerful and general learning and reasoning capability, using a powerful general purpose representational formalism (e.g. logic, semantic nets) can bootstrap everything else. But so far all such attempts have fallen foul of combinatorial problems. (GPS, theorem provers.)

- Rationalist AI, based on the assumption that anything that can achieve what a human does in a lifetime must start with a massive amount of prior knowledge, has also not achieved much.
- But that's partly because nobody has tried to work out in a systematic way what sorts of apriori information-processing capabilities are required.
- McCarthy's paper illustrates some of the sort of things we have to specify and explain. But he does not say much about the architecture required.
- I suspect we don't yet understand the nature of the problem. The variety of different innate learning capabilities in different organisms may provide useful clues (e.g. ask: what can't a cat learn and why not). Also genetic brain deficiencies.
- We'll also have to ask about different innate learning capabilities required in different parts of the same multi-layered organism or robot.

Meanings in architectures

The ‘CogAff’ Schema

In a multi-faceted architecture there may be different kinds of meanings/concepts/information structures at work in different sub-architectures

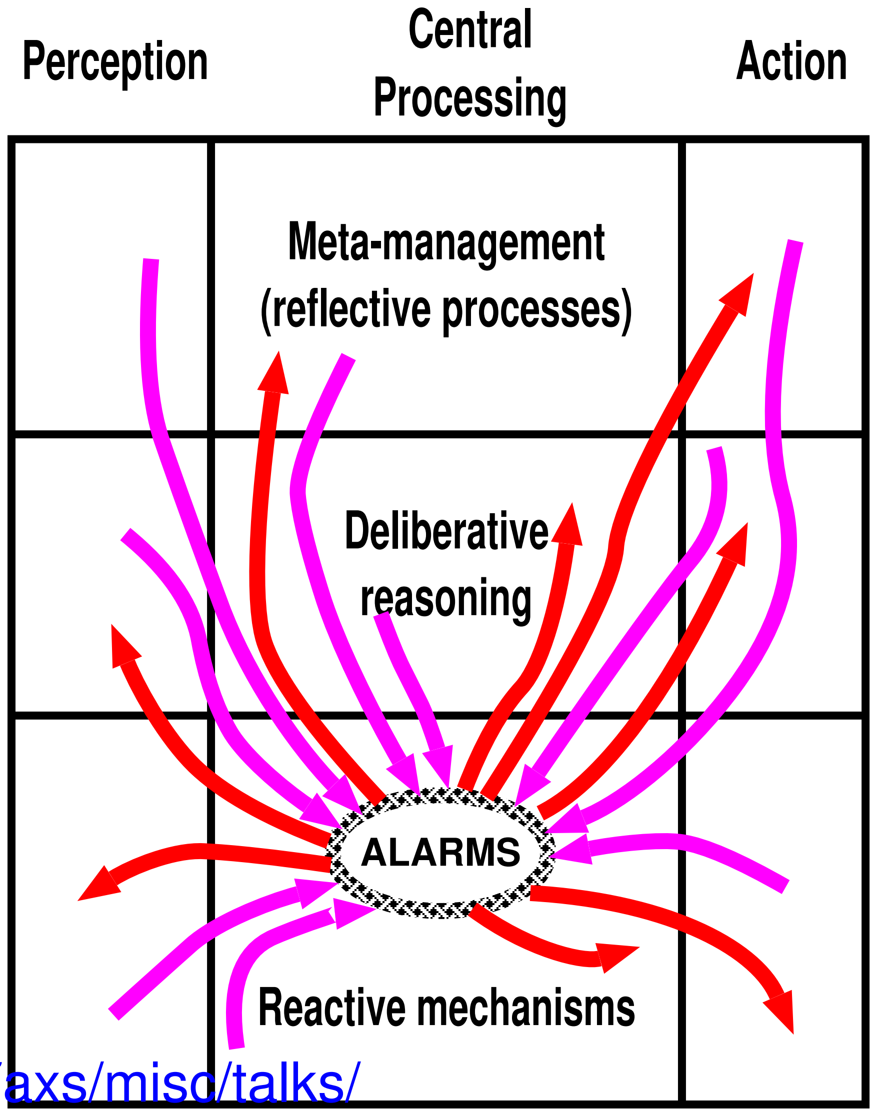
Different kinds of functionality are represented by different boxes in the CogAff schema.

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

See http://www.cs.bham.ac.uk/˜axs/misc/talks/
See Minsky’s Society of Mind

An alarm mechanism may have its meanings

If a trainable alarm mechanism uses self-organising neural nets, the kinds of meanings it finds in the patterns that provoke global interrupts may be very different from the kinds of meanings used in other parts of the system.



See http://www.cs.bham.ac.uk/˜axs/misc/talks/

Different kinds of apriori “concepts” in different parts of an architecture?

A previous slide stated:

From the previous conjectures it follows that grasping colour concepts may simply be a matter of plugging certain parameters (perhaps procedural parameters) into a complex mechanism — where alternative parameters are required for other concepts of properties of extended surfaces, e.g. texture, warmth, etc.

If a new-born, or newly hatched, or newly manufactured, animal or robot has an architecture with the sort of variety of components indicated previously, then for each of the components to work it will need to process information.

Perhaps the innate mechanisms in different layers, and the concepts they support, will vary:

E.g. if there’s a reactive layer and a deliberative layer from the start, the mechanisms supporting spatial information in the two layers may be quite different. Or maybe the same general mechanism is used, but with different parameters for the different layers? Or perhaps things are different for the different sub-mechanisms within the layers?

(Minsky: The society of mind)

A related issue: Strong AI vs Weak AI thesis

Searle made this distinction in the late 1970s.

- The weak AI thesis: It is possible (in principle, in the future) to build computer-based machines that behave *as if* they understand symbols, have mental processes, etc. But no matter how good the simulations, there will not be replication of mentality, intentionality, understanding, feeling, etc. (Compare simulations of tornadoes.)

  Note: such machines may have “derivative intentionality” insofar as WE interpret their internal symbols and external utterances as meaningful.

- The strong AI thesis: It is possible (in principle, in the future) to build (computer-based) machines that do not merely behave *as if* they understand symbols, have mental processes, etc. but also *really do* understand etc. I.e. there will be not just *simulation* but actual replication of mentality, intentionality, understanding, feeling, etc.

  Note: such machines will have “original (non-derivative) intentionality” insofar as THEY understand their internal symbols and external utterances.

The Strong AI thesis is inherently ambiguous (e.g. what’s “computer-based”):

On some interpretations it is obviously false, on some it is obviously true, and some pose open research problems.

See my 1992 review of Penrose, in AIJ, on the CogAff web site.

Evolution of information-processing virtual machines

Evolution "discovered" and used many things long before human engineers and scientists thought of them.

Paleontology shows the development of physiology and provides some weak evidence about behavioural capabilities.

But there is very little direct evidence regarding previous forms of information processing: virtual machines leave no fossils.

Forms of information processing now found in nature give clues, and we can test theories in working models.

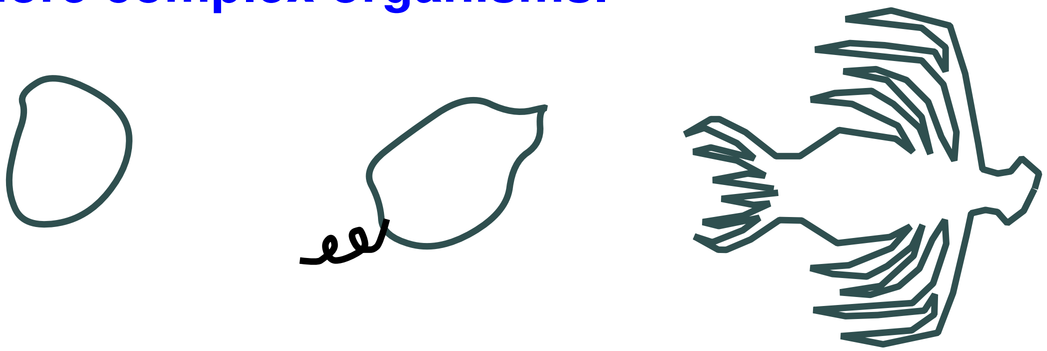
Some of the forms are evolutionarily very old. Others relatively new. (E.g. the ability to learn to read, design machinery, do mathematics, or think about your thought processes.)

WE NEED TO LEARN HOW TO ASK NEW (DEEP) QUESTIONS ABOUT THE POWERS AND FUNCTIONS OF DIFFERENT SYSTEMS, AND HOW THEY FIT TOGETHER.

Organisms process information

Once upon a time there were only inorganic things: atoms, molecules, rocks, planets, stars, etc.
These merely reacted to *resultants* of all the physical forces acting on them.

Later, there were simple organisms. And then more and more complex organisms.



These organisms had the ability to reproduce. More interesting was their ability to *initiate* action, and to *select* responses, instead of simply being pushed around by physical forces acting on them.

That achievement required the ability to acquire, process, and use *information.*

We use “information” in the everyday sense, not the Shannon/Weaver technical sense

Acting or selecting requires information

E.g. information about

- density gradients of nutrients in the primaeval soup
- the presence of noxious entities
- where the gap is in a barrier
- precise locations of branches in a tree as you fly through
- how much of your nest you have built so far
- which part of the nest should be extended next
- where a potential mate is
- something that might eat you
- something you might eat
- what that thing over there is likely to do next
- how to achieve or avoid various states
- how you thought about that last problem
- whether your thinking is making progress

and much, much more... (has anyone attempted a taxonomy?)

Most of these processes don't involve self-consciousness.

Resist the urge to ask for a definition of "information"

Compare "energy" – the concept has grown much since the time of Newton. Did he understand what energy is?

Instead of defining "information" we need to analyse the following:

- the variety of types of information there are,
- the kinds of forms they can take,
- the means of acquiring information,
- the means of manipulating information,
- the means of storing information,
- the means of communicating information,
- the purposes for which information can be used,
- the variety of ways of using information.

As we learn more about such things, our concept of "information" grows deeper and richer.

Like many deep concepts in science, it is *implicitly* defined by its role in our theories and our designs for working systems.

Things you can do with information

A partial analysis to illustrate the above:

- You can react immediately (it can trigger immediate action, either external or internal)
- You can do segmenting, clustering labelling of components within a complex information structure (i.e. do parsing)
- You can try to derive new information from it (e.g. what caused this? what else is there? what might happen next? can I benefit from this?)
- You can store it for future use (and possibly modify it later)
- You can consider alternative possibilities, e.g. in planning.
- If you can interpret it as as containing instructions, you can obey them, e.g. carrying out a plan.
- You can observe the process of doing all the above and derive new information from it (self-monitoring, meta-management).
- You can communicate it to others (or to yourself later)
- You can check it for consistency, either internal or external

... using different forms of representation for different purposes.

Intentionality and semantics

Intentionality involves the ability to refer to something (in thoughts, desires, plans, explanations, questions, etc.) I.e. it involves semantics (meaning).

Often assumed to be a requirement for consciousness: consciousness is always OF something.

John Haugeland distinguished derivative and original intentionality. Printed words, maps, footprints, etc. have derivative intentionality: they can only be used by something else to refer.

Philosophers often write as if “original intentionality” is either present or absent.

By exploring the variety of architectures for machines that process information we can distinguish a wide range of cases: varieties of intentionality.

Likewise if we look at many natural phenomena, e.g. people with brain damage, newborn infants, different sorts of animals.

A new distinction between architecture-based and architecture-driven concepts

Key idea: certain concepts (forms of information content) arise out of the fact that a system has a certain sort of information processing architecture, which does various things. (E.g. the concept of ‘qualia’, the concept of ‘infinity’.)

This will be explained later.

WATCH THIS SPACE.

People designing intelligent robots

will ultimately find all this more useful than symbol-grounding theory.
Likewise people working on automated learning and ontology formation.

Old problems for concept empiricism: Logic and Mathematics

How could logical concepts be based on abstraction from instances? What would experiencing instances of "not" and "or" and "if" be like? (At least one philosopher thought understanding "or" might be based on experiences of hesitancy.)

Concepts of number were more debatable: Mill thought they all came from experience. Frege thought they could be defined using pure logic?

(.... his "proof" fell foul of Russell's paradox)

Perhaps they are both right and they are multi-faceted concepts?

What about our concept of infinity? Can any instances of infinity be experienced? Can we define it adequately in terms of a concept of negation and "being bounded"? Does that suffice to support our thinking about transfinite ordinals (e.g. arranging the odd integers after the even ones)?

What about concepts of euclidean geometry: infinitely thin, infinitely straight, infinitely long lines? Think of the continuum.

Many mathematical concepts are purely structural concepts, capable of being applied to wide ranges of portions of reality, e.g. the concept of a group, a permutation, a function, an ordering.

Could mere experience of instances generate the apparatus required for grasping and using such generally applicable concepts?

More problems for concept empiricism: Concepts of theoretical science

Deep scientific theories refer to things that cannot be experienced, e.g. gravitational fields, electrons, nuclear forces, genes, ...

Some philosophers tried to argue that the concepts of theoretical science can all be defined in terms of observables (e.g. things that can be measured, where the measuring process is experienced). All attempts to *define* theoretical concepts like this broke down, e.g. because the modes of measurement seemed not to be definitional – they could change without the concepts changing.

Example: Bridgman's "operationalism" (1927)

There was also the question about the existence of theoretical entities, states, processes while they were not being measured or observed.

Dispositional properties (solubility, fragility, rigidity, etc.) also raised problems.

How can we understand the notion of objects having dispositions (e.g. solubility) and capabilities (e.g strength) while they are not displaying them, so that they are not experienced?

How can we understand the concept of existence of something unexperienced (the tree in the quad when nobody's there)?

(Some philosophers said we don't.)

Counterfactual conditionals

Various attempts were made to cope with this via definitions in terms of large collections of counter-factual conditionals.

“There is an electron with properties P, Q, R, ...”

means

“if you do such and such experiments then you will observe such and such results...” etc.

But no finite collection of conditions seemed to be capable of exhausting any such concept, e.g. because new (and better) experimental tests could be discovered, and old tests rejected.

Some philosophers (phenomenalists, e.g. Hume) tried this for all concepts referring to things that exist independently of us, since they can have properties that are not manifested in our experience, if we are asleep or not looking at them, etc. (Objects as permanent possibilities of experience.)

How can you have a concept of something that exists while you are not experiencing it? Or of an unexperienced possibility. Can you experience that?

The theory that God experiences all those things continuously, including the tree in the empty quad does not help, especially atheist philosophers.

Also where does the concept of God come from? (Software bugs in minds)

Harnad on symbol systems

According to his 1990 paper, a symbol system is:

1. a set of arbitrary “physical tokens” scratches on paper, holes on a tape, events in a digital computer, etc. that are
2. manipulated on the basis of “explicit rules” that are
3. likewise physical tokens and strings of tokens. The rule-governed symbol-token manipulation is based
4. purely on the shape of the symbol tokens (not their “meaning”), i.e., it is purely syntactic, and consists of
5. “rulefully combining” and recombining symbol tokens. There are
6. primitive atomic symbol tokens and
7. composite symbol-token strings. The entire system and all its parts – the atomic tokens, the composite tokens, the syntactic manipulations both actual and possible and the rules – are all
8. “semantically interpretable:” The syntax can be systematically assigned a meaning e.g., as standing for objects, as describing states of affairs).

Some problems with symbol grounding theory.

Objection:

the symbols do not need to be physical symbols. They can be abstract entities in a virtual machine, e.g. lists, graphs, trees, in Lisp.

Objection:

his account does not say anything about what the system *does* with the symbols.

Objection:

arbitrariness is a red herring, though often cited by philosophers:. They forget that symbols normally form systems with *systematic* modes of composition.

Objection:

in his grounding theory he requires primitive symbols to be structurally isomorphic with things in the environment that produce them. That arbitrarily rules out potentially useful designs. E.g. must smell sensors be like that?

Objection:

Symbols do not “stand for” the things they refer to. Symbols and the things they refer to are used for quite different purposes. Do not think of the word “table” as a substitute for a table.

[[This page is to be revised]]

Confusions found in symbol-grounding theories

- If structures are merely created, rearranged, destroyed, to no purpose, they are not being used as symbols — in discussing symbol systems we need to discuss the architecture in which they are *used*.
  Compare patterns forming and changing when leaves are blown about by the wind.
- If they are being used either to control processing, or as intermediate states of processing in a system that achieves something, e.g. compiling a program, time-sharing a computer, managing a filing system, garbage collecting virtual memory, driving perceptual mechanisms to ask questions about the environment, operating on perceptual data so as to find features, analyse structures, interpret structures, then in part what they mean will be determined by that role.

  Examples: processes in an operating system, such as scheduling, file management, memory management, access checking.
- Some of that role (e.g. the role of structures and mechanisms in a perceptual system) may be defined largely in terms of what happens within the system: if the mechanisms operate on structures produced by interacting with an environment, that interaction *extends* the role of the symbols.

More “potted” history of philosophy of AI

Many have asked how it is possible for machines to *understand* symbols they manipulate internally, or symbols they emit or read in.

‘Can machines think?’ is an old question discussed by philosophers since the earliest days of computers (and before that: e.g. in the writings of Ada Lovelace).

Turing posed the question in his 1950 article, but was too intelligent to discuss it!

He noticed that the question was far too vague to have an answer. Instead he made a technological prediction – machines will be able to pass his test, widely misconstrued as a criterion for thinking or intelligence, which it never was.

Dennett, Haugeland and Searle discussed the question in the 1970s and 1980s.

I wrote some papers on it e.g. in IJCAI-85 ECAI-86 (both on Cogaff web site)

Symbol grounding

In the late 1980s Stevan Harnad started writing about “the symbol grounding problem” and putting forward his answer, e.g.

http://cogprints.soton.ac.uk/documents/disk0/00/00/06/15/
The Symbol Grounding Problem. Physica D 42:335-346. (1990)

Newcomers to AI and Cognitive science often regard his theory almost as axiomatic.

Compare Gardenfors on “Conceptual Spaces”.

Original (intrinsic) intentionality

From this standpoint the assumption that there's a clear division between things with and things without "original intentionality" is mistaken.

Neither is it just a matter of differences of degree.

There's a discrete set of information-manipulating capabilities and different subsets can coexist. No subset corresponds uniquely to any pre-theoretical notion of "understanding", or "intentionality".
(I.e. we are dealing with "cluster concepts". That claim needs argument.)

Of course, different theorists can decide to attach the label "intentional" to some particular subset, but that's far less interesting than trying to understand the full range of possibilities.

So instead of asking whether Z does or does not have original intentionality we should ask "What sort of intentionality does Z have?

We can then produce a systematic taxonomy of cases, investigate their similarities and differences, and learn something instead of simply supporting our prejudices.

Meaning postulates

Rudolf Carnap introduced the idea of 'meaning postulates'.

Search for that phrase in this document:
http://www.utm.edu/research/iep/c/carnap.htm

Or read his paper on meaning postulates in his 1947 book: Meaning and Necessity.

Carnap's primary motivation for introducing meaning postulates was to explain how dispositional concepts and theoretical concepts in the advanced sciences can be understood, but the idea is far more general than that.

What I've written above about the structure of a set of representations and mechanisms for using them partially defining the "meaning" of the primitive components is inspired by Carnap's theory of meaning postulates, which can introduce new undefined primitives into a theory in the form of new postulates (axioms) using the new primitives alongside the old ones.

The meaning can be gradually refined and made more determinate by adding more postulates. We can say it's a new meaning after each change. Or we can treat it like a river and say it's the same meaning which gradually changes over time.

I don't know if Carnap would approve of my use of his ideas.

Acknowledgements

This research is supported by a grant from the Leverhulme Trust.

I have had much help and useful criticism from colleagues at Birmingham and elsewhere, especially Matthias Scheutz and Ron Chrisley.

The work is informal and partly sketchy and conjectural rather than a polished argument, even though I have thinking about the problems on and off for about 30 years. Comments and criticisms welcome.

Some references

Some of the ideas are in these two old papers:

`http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#4`
"What enables a machine to understand?" (IJCAI 1985)

`http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#5`
"Reference without causal links" (ECAI 1986)

and in Chapter 2 on the aims of science, in *The Computer Revolution in Philosophy*, now available free of charge online:

`http://www.cs.bham.ac.uk/research/cogaff/crp/chap2.html`

The Birmingham Cognition and Affect Project

PAPERS (mostly postscript and PDF):

`http://www.cs.bham.ac.uk/research/cogaff/`
(References to other work can be found in papers in this directory)

TOOLS:

`http://www.cs.bham.ac.uk/research/poplog/freepoplog.html`
(Including the SIM_AGENT toolkit)

SLIDES FOR TALKS (Including IJCAI01 philosophy of AI tutorial with Matthias Scheutz):

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

Comments, criticisms and suggestions for improvements are always welcome.