Why symbol-grounding is both impossible and unnecessary, and why symbol-tethering based on theory-tethering is more powerful anyway.

Introduction to key ideas of semantic models,
implicit definitions and symbol tethering*

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

With help from Ron Chrisley, Jackie Chappell and Peter Coxhead

Substantially modified/extended in June 2008

[*] Symbol-tethering was previously described as Symbol-attachment

These slides will be available in my "talks" directory at:
`http://www.cs.bham.ac.uk/research/cogaff/talks/#models`

This is an expanded version of part of an older set of slides criticising
Concept Empiricism and Symbol Grounding theory
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#grounding`

Apologies

I apologise

For slides that are too cluttered: I write my slides so that they can be read by people who did not attend the presentation.

So please ignore what's on the screen unless I draw attention to something.

NO APOLOGIES for using linux and latex instead of powerpoint.

Acknowledgement:

This work owes a great deal to Kant, to logicians and mathematicians of the 19th and 20th century, and to the logical and philosophical work of Tarski, Carnap, and other 20th century philosophers of science.

However I believe all that work is merely preliminary to what still needs to be done, namely showing how to build a working system that can extend its concepts in the process of learning about the world and discovering that there is much more to the world than meets the eye or skin.

I suspect that making progress on that task will require us to discover new forms of representation, new mechanisms for operating on them and new information-processing architectures combining them – new to science, engineering and philosophy, even if they are very old in biological systems.

Some of the ideas presented here were in chapter 2 of *The Computer Revolution in Philosophy: Philosophy science and models of mind* (1978) `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`
What are the aims of science?
First published in *Radical Philosophy* (1976)

Two sorts of empiricism

Concept empiricism is an old, very tempting, and mistaken theory.
(Recently re-invented as "symbol-grounding" theory (S-G theory below).)

It is related to, but different from, **knowledge empiricism**.

One of the most distinguished proponents of both sorts of empiricism was the philosopher David Hume, though the ideas are much older.

Immanuel Kant argued against both kinds of empiricism.

Knowledge empiricism states:

> All knowledge (about what is and is not true) has to be derived from and testable by sensory experience, possibly aided by experiments and scientific instruments.

It usually allows for mathematical and logical knowledge as exceptions, though sometimes describing such knowledge as essentially empty of content (like definitional truths).

> Hume's knowledge empiricism:
>
> "When we run over libraries, persuaded of these principles, what havoc must we make? If we take in our hand any volume of divinity or school metaphysics, for instance, let us ask, Does it contain any abstract reasoning concerning quantity or number? No. Does it contain any experimental reasoning concerning matter of fact and existence? No. Commit it then to the flames, for it can contain nothing but sophistry and illusion."
>
> *An Inquiry Concerning Human Understanding*

Concept Empiricism is about Concepts

Concept empiricism is a theory about concepts, the building blocks required for both true and false judgements, beliefs, theories, hypotheses, ...

It is not always noticed that concepts are also required as building blocks for all sorts of mental contents, including questions, desires, intentions, plans, preferences, hopes, fears, dreams, puzzles, explanations, etc.

Concepts can be thought of as the building blocks of all semantic contents.

They are also required as building blocks for non-conscious contents, e.g. subconscious intermediate states in perceptual processing or motor control, and stored factual and procedural information used in understanding and generating language, controlling actions, learning, taking decisions, etc.

Some people think there are also non-conceptual contents, though what that means is not clear, and will not be discussed here.

If an animal or machine makes use, consciously or otherwise, of semantic contents using a form of representation that allows novel meanings to be expressed by construction of novel meaning-bearers, then whatever the components are that can be rearranged or recombined to produce such novelty can be called “conceptual”. Concepts corresponding to properties and relations are a special case.

There are also concepts that are *potentially* available for use even though no animal or machine has ever used them - e.g. a concept of a type of animal with a large number N of heads, for some N which I am not going to specify. Perhaps one of those concepts will occur in a future science fiction story.

Concepts in this context are not primarily for communication

Concept empiricism is not just a theory about conditions for understanding concepts used in communication, but about concepts needed for thinking, wanting, supposing, intending, wondering whether, wondering why, etc.

Likewise the symbols (representations) we'll be talking about are not primarily symbols in an external language, but the ones needed by an animal or machine for internal use.

It is often thought that concepts, the building blocks of thoughts, questions, gaols, plans, etc. must always be expressed in discrete symbolic structures that occur in a language with a formal syntax and compositional semantics, that is used for communication.

I have argued elsewhere that we need to generalise that notion to include "generalised" languages (GLs) used only for *internal purposes*, some of which may be more pictorial or spatio-temporal than symbolic formalisms.

See the slide below, entitled *"Generalising the notion of 'language' or 'formalism'"*, and also
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`
What evolved first: Languages for communicating, or languages for thinking (Generalised Languages: GLs)

GLs of some form are needed to explain some of the intelligent behaviours of pre-verbal children and non-verbal animals.

However, for much of the rest of this presentation I shall deal only with linguistic or logical formalisms, where some of the issues are clearer – though I regard this as merely preliminary clarification which must later be extended to GLs including analogical representations.

What alternatives are there to Concept Empiricism?

It is often assumed that the only alternative to concepts being derived by abstraction from experience of instances is that they must be innate.

Compare: J.A. Fodor *The Language of Thought* (1975)

But the history of science shows that many concepts have been introduced by research communities that are not explicitly definable in terms of concepts that were used earlier.

Examples include concepts like “gene”, “electron”, “charge”, “valence”, and many more.

So concept empiricism (S-G theory) must be false.

Likewise a child of ten uses many concepts that are not definable in terms of the conceptual apparatus of a child of three months – yet the later concepts must in some way grow out of the activities of the child, use previously available competences.

There is some sort of boot-strapping process that needs to be explained.

The rest of this presentation aims to show how the introduction of new concepts can go hand in hand with the introduction of new *theories* that not only use those concepts, but also implicitly, and partially, define them.

The process by which a child or scientist (or future robot) develops substantively new concepts depends on the processes that significantly extend her theories of how the world works, simultaneously creating and making use of those new concepts:

New theories don’t simply express laws relating old ideas: New ideas are developed also.

Where logic is used, the process can simultaneously involve **abduction** (adding new premisses to explain old facts) and **adding new symbols, not definable in terms of old ones.**

If concept empiricism (S-G theory) were true....

It is often assumed (ignoring some the power of biological evolution) that a new-born animal has access **only** to concepts expressible in terms of what it can experience, or more generally concepts definable in terms of patterns in its sensory and motor signals.

In combination with concept empiricism (or symbol-grounding theory) that assumption implies that ALL concepts ever used are restricted to describing patterns and relationships (e.g. conditional probability relationships) in sensory and motor signals.

Since sensory and motor signals are processes occurring within the body, an ontology referring to patterns and relationships in those signals could be called a "somatic" ontology (from the Greek word "soma" meaning "body").

So concept empiricism implies that no animal or machine can ever think about, refer to, perceive, have beliefs, or have goals that concern anything that happens or could happen outside it: i.e. using exo-somatic concepts would be impossible.

If visual sensory inputs form a changing 2-D array, then, according to concept empiricism, a visual system could not provide information about 3-D structures, and a perceiver with only visual senses could never even conceive of 3-D structures: the Necker cube ambiguity would be impossible.

Compare Plato's cave-dwellers perceiving only shadows on the wall of the cave: concept empiricism would prevent them thinking of them as shadows of something else, different in kind.

It would also be impossible to think of space or time as extending indefinitely, or of events that occurred before one's birth, or of objects composed of unobservable particles with unobservable properties.

Two kinds of ontology extension

An important aspect of learning and development is acquiring new concepts, providing the ability to have new mental contents, referring to new kinds of entity: Ontology extension.

Ontology extension may be

- **definitional**

  New concepts added that are explicitly defined in terms of old ones: this is essentially just abbreviation, though often very useful.

- **substantive.**

  New concepts added that cannot be explictly defined in terms of old ones.

Substantive ontology extension is an important aspect of scientific advance, contrary to the view that science is merely concerned with discovering new facts, e.g. laws of nature.

See Sloman 1978, Chapter 2. "What are the aims of science?"
`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`

If all concepts had to be explicitly defined in terms of others, that would produce either circularity or an infinite regress.

So either some concepts must be innate (or available when a machine first starts up) or there must be a way of acquiring new concepts that does not rely on definition in terms of old concepts.

According to concept empiricism the only way to acquire new concepts that are not definable (i.e. the only method of substantive ontology extension) is:

by abstraction from experience of instances.

Exo-somatic concepts can be innate

The notion of substantive, non-definitional, extension of concepts is not restricted to concepts acquired in a process of learning and development.

It is possible for concepts that are provided by genetic mechanisms to form different subsets, some of which are substantive extensions of others.

For example, suppose a new-born animal, or robot, has two sorts of concepts from birth, as a result of evolutionary processes:

- Somatic concepts – referring only to patterns in sensor and motor signals and relationships between them, including probabilistic and conditional relations, e.g.:

  A concept of sensed temperature, i.e. feeling hot or cold; or the concept of being in a shivering state, defined by fast rhythmic signals being sent to certain effectors; or the dispositional concept of being in a state of feeling cold that will change to a state of feeling hot if shivering happens.

- Exo-somatic concepts – referring to things, states of affairs, events, and processes that exist outside the organism, and whose existence at any time has nothing to do with whether they happen to be sensed or acted on at that or any other time, e.g.:

  a concept of some bushes existing nearby; a concept of a dangerous predator being hidden in the bushes, and a concept of the predator attempting to eat another animal in the future.
  A migrating bird might innately have a concept of a distant location to which it should go.
  (In some species, young birds first migrate after their parents have gone.)

In such cases, the exo-somatic concepts are substantive additions to the somatic concepts, even though both sets are innate.

But exo-somatic concepts can also be developed by an individual, as we'll see.

(Also by a scientific culture.)

Concepts and Symbols/Representations

Concepts, and the larger structures built using them, are usable only if expressed or encoded in something.

This requires a medium and mechanisms to use the medium, e.g. manipulating information structures.

The words “symbol” and “representation” refer to the contents of the medium.

Many cases can be distinguished

Some media support only fixed structures, e.g. certain switches that can be on or off, or fixed-size vectors of values.

Others allow new structures to be created

Human percepts and thoughts require a medium in which new structures can be created of varying complexity.

> Compare: seeing one black dot on a large white wall, seeing a busy city-centre street, seeing ocean waves breaking on rocks.

Examples of representing structures include: vectors, arrays, trees, graphs, maps, diagrams, dynamic simulations, ...

NOTE: The relationship between symbols (representational structures) and what they express (concepts, thoughts, ...) can be complex, subtle and highly context-sensitive: there need not be a 1-to-1 mapping. See Sloman 1971, here (also chapter 8 of Sloman 1978):

`http://www.cs.bham.ac.uk/research/cogaff/04.html#200407`

Symbols can exist in virtual machines

Symbols need not be physical: they can be virtual structures in virtual machines, and usually are in AI systems.

Newell and Simon (amazingly) got this wrong in their "physical symbol system" hypothesis.

In almost all AI systems the symbols are not physical but occur in **virtual** machines, e.g. Lisp or Prolog atoms, or similar entities in other languages.

NOTE
In what follows I may slide loosely between talking about symbols and talking about the concepts they express.

E.g. I talk both about somatic and exo-somatic concepts and about somatic and exo-somatic symbols or representations.

In the latter case it is the semantic content of the symbols or representations that is exo-somatic, even when the symbols or representations themselves (the meaning-bearing structures) are within the body of the user.

I hope the context will make clear which I am referring to.

Thanks to Peter Coxhead for pointing out the need to be clear about this.

Note: intelligent agents are not restricted to using internal symbols and representations. You can reason with diagrams in your mind or on paper.

Recap: Concept empiricism states (roughly):

All concepts are ultimately derived from experience of instances

- All simple concepts have to be abstracted directly from experience of instances
- All non-simple (i.e. complex) concepts are constructed from simple concepts using logical and mathematical methods of composition.

  "When we entertain, therefore, any suspicion that a philosophical term is employed without any meaning or idea (as is but too frequent), we need but enquire, from what impression is that supposed idea derived? And if it be impossible to assign any, this will serve to confirm our suspicion."
  David Hume, *An Enquiry Concerning Human Understanding* Section II.

  E.g.

  if red and line are simple concepts
  then logical conjunction can be used to construct from them
  the concept of a red line.

  Hume used the example of the concept of a golden mountain.

  Hume also allowed a mode of invention of a concept by *interpolation*,
  e.g. imagining a colour never before experienced, by noticing a gap in the ordering of previously experienced colours.

  Mathematical logic allows many more ways of defining new concepts from old than I suspect Hume ever dreamt of.

  Programming languages extend that: complex predicates can be defined in terms of complex procedures for testing instances.

  Mutual definition is also possible in mathematics and programming languages (mutual recursion).

Symbol Grounding Theory

Symbol grounding theory is a modern (re-invented) version of concept empiricism.

The term “symbol grounding” was introduced by Stevan Harnad in 1990.

Stevan Harnad, The Symbol Grounding Problem, *Physica D* 42, pp. 335–346,
`http://users.ecs.soton.ac.uk/harnad/Papers/Harnad/harnad90.sgproblem.html`

Symbol grounding theory adds extra requirements to concept empiricism that I shall ignore e.g.

the experience of instances must use sensors that provide information in a structure that is close to the structure of the things sensed

The label “symbol grounding” has proved to be a very powerful viral meme, and has damaged many philosophically uneducated minds, and some of the educated ones also.

E.g. according to this and other versions of concept empiricism, theoretical concepts of science, that have nothing to do with sensorimotor signals of scientists, and, more generally, exo-somatic concepts are impossible.

Roboticists and AI researchers who accept symbol-grounding theory restrict the forms of learning in intelligent machines in seriously damaging ways:

- They cannot learn exo-somatic concepts, only somatic concepts.
- They use ontologies restricted to patterns and relationships in their sensor and motor signals.
- Often they ignore the structure of the environment and assume everything to be learnt can be derived from arbitrary arrays of bit-patterns in sensors. (Evolution would not make such a mistake.)

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 prior concepts and some formal apparatus for expressing definitions.)
- OR You have experience of examples, and then you somehow get the idea.

(How you get the idea is usually left unexplained by philosophers, but AI theorists have produced a number of working mechanisms.)

Two sub-cases of the second case:

- Substantive concept formation What you learn is not definable in terms of prior concepts using any formalism you can articulate
  E.g. learning to recognise a colour as a result of seeing examples, or learning the taste of kiwi fruit from examples.
  (This might be implemented using chunking in self-organising neural nets.)
- Concept formation by explicit definition
  An explicit definition of the new concept is extracted from the examples.
  E.g. Winston's 1970 program learning about arches, houses, etc., made of simple blocks. — amounts to guessing an explicit definition.
  Many AI concept-learning programs learn only concepts that are explicitly defined in terms of pre-existing concepts.
  So all the concepts they learn are definable in terms of some initial primitive set of concepts.
  They do not do "substantive" concept formation.

Concept empiricism seems irrefutable, at first

People are tempted by concept empiricism (or symbol grounding theory) because they cannot imagine any way of coming to understand certain notions except by experiencing instances.

E.g.

red, sweet, pain, pleasure, curved, bigger, straight, inside, ...

etc.

They then conclude that this is the only way of acquiring new concepts, apart from explicit definition of new concepts in terms of old concepts.

What sort of theory is concept empiricism?

We have no introspective insight into the processes that produce new concepts by abstraction from experiences.

People usually assume that we can somehow store something unexplained that is abstracted in some unexplained way 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 a prototype for each concept and compare it with new instances",
etc., etc.

Trying to go from such a verbal description to a specification of a working system that can be built and run often shows that the alleged explanation is not capable of explaining anything, or that there are many different detailed mechanisms that loosely correspond to the verbal description because it is too vague to select between them.

More people should learn how to use the design stance.

If you have first hand experience of designing (and debugging) systems that have to work, you may

(a) be able to think of a wider variety of possible explanations for observed phenomena,
(b) be better at detecting theories that could not possibly work,
(c) be better at detecting verbal theories (with or without accompanying diagrams) that are so intrinsically vague that there many very different working systems that would fit the description given.

Kant's refutation of concept empiricism

Over two centuries ago, the philosopher Immanuel Kant refuted concept empiricism

(in his *Critique of Pure Reason*, 1781).

His main argument (as I understand it) can be summarised thus:

> It is impossible for all concepts to be derived ultimately from experience, since if you have no concepts you cannot have experiences with sufficient internal structure to be the basis of concept formation.
>
> So some concepts must exist prior to experience.
>
> He gives different arguments for some specific concepts that he thinks must be innate, such as concepts of time and causation.

Let's consider an example of a spatial concept.

An example: straightness

According to concept empiricism there are two ways to acquire a concept such as straightness

Either

- It is an unanalysable concept learnt from experience of examples.

or

- It is a **composite** concept explicitly defined in terms of some previously understood concepts.

Case 1: straightness is unanalysable.

The notion that you can use experiences of straightness to abstract the concept "straight" presupposes that you can have spatial experiences (e.g. of straight lines, straight edges, straight rows of dots, etc.) without having concepts — usually that presupposition is not analysed.

What makes something straight is having parts with certain relationships, so detecting what is common to examples of straightness requires the ability to detect portions of what is seen and relationships between them. This does not seem to be possible without using concepts.

Case 2: straightness is a definable composite concept

An objection would be that "straight" is a composite concept that has to be explicitly defined in terms of simpler concepts, so it is only those simpler concepts (including concepts of relationships) that need to be learnt from experience of instance.

For this to be taken seriously, an explanation is required of what those simpler unanalysable concepts are and how they can be learnt and used without using concepts.

Two things wrong with Concept Empiricism

- The first flaw in concept empiricism was pointed out clearly by Kant as explained above.

  YOU CAN'T HAVE EXPERIENCES UNLESS YOU ALREADY HAVE CONCEPTS,
  SO NOT ALL CONCEPTS CAN BE DERIVED FROM EXPERIENCES OF INSTANCES.

- The second flaw was that supporters of concept empiricism ignored problems about new concepts introduced in the development of science.

  These were discussed in the 20th century by philosophers of science.

  They failed to think of an alternative way in which concepts can be developed which is illustrated by theoretical concepts in Science.

  20th Century philosophers of science showed that such theoretical concepts (e.g. in physics) cannot be defined in terms of how instances are experienced, and cannot be defined operationally (as proposed by P.W.Bridgman).

This included trying to explain how new undefined theoretical concepts could work.

- Concepts of unobserved entities and properties, i.e. theoretical concepts in scientific theories
  (e.g. the distant past, the properties of matter that explain how matter behaves)
- Concepts of dispositions that are not manifested
  (e.g. solubility, fragility, being poisonous, etc.)

There are also problems about how logical and mathematical concepts that do not refer to specific sorts of experiences can be understood (e.g. "not", "all", "or", "number", "multiplication"), and concepts like "efficient", "useful", "uncle", "better"....

Symbol grounding theory and its alternatives

Harnad wrote in his 1990 paper:

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

We are presenting an alternative suggestion, based on 20th Century philosophy of science, which provided an account of how theoretical concepts in science work (e.g. 'electron', 'charge', 'ion', 'chemical valence', 'gene', 'inflation'):

- Many kinds of meaning are constructed in an abstract form, as part of a theory that uses them:

  new symbols may be added to a theory in theoretical statements that use them, and the meanings are partly determined by the structure of the theory and the forms of inference that can be made.

- This always leaves meanings partly indeterminate.
- They can be made more definite in two ways
  1. by making the theory more complex, adding new constraints to the interpretations of undefined symbols
  2. by linking the theory to methods of observation, measurement and experiment.

I.e. existing, un-grounded meanings are made more definite as the theory using them becomes "tethered" – but tethering does not define the symbols in the theory.

The KEY new idea: implicit definition by a theory

The key new idea in the alternative philosophy of science, was developed independently by Einstein and by philosophers of science:

(See `http://plato.stanford.edu/entries/einstein-philscience/`),

If a new explanatory theory with predictive power, introduces some new theoretical concepts, those concepts can get most of their meaning from their role in the theory: they are implicitly defined by the theory containing them.

To explain this we need to introduce the idea of a theory (including undefined symbols) as determining a class of models.

This is not a new suggestion: the idea was developed by 20th Century philosophers of science.

I discussed it in Chapter 2 of Sloman 1978, and in these two papers:

What enables a machine to understand?, *Proc 9th IJCAI, 1985*
`http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#4`

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

Why tethering is needed

A theory need not uniquely determine a model – so it may need to be "tethered" by adding associations with observation and experiment.

(Thanks to Jackie Chappell for suggesting the word "tethering".)

We shall shortly present an example of a theory, using a logical syntax, including

- conjunction,
- disjunction,
- negation,
- universal and existential quantification,
- application of predicates to arguments, and
- application of functions to arguments.

We show how elaborating the theory reduces the set of models, but something else is needed to refer to a specific portion of the world.

Introducing new theoretical concepts

We need to explain how a formal theory, or set of axioms can produce the first, un-tethered set of meanings.

A Key Idea: A formal System

KEY IDEA

A formal system with a well defined syntax and rules of inference allows a set of 'axioms' possibly including undefined predicate and function symbols, to determine a class of possible interpretations (models).

The set of axioms gives the undefined symbols a (partially indeterminate) meaning.

Example on next slide: axioms for lines and points (projective geometry)

- A formal system specifies syntactic forms for constructing new meanings from old, and forms of inference for deriving new conclusions from old hypotheses.
- We assume that those are familiar ideas in Logic and will use logically formulated theories as our example.
- But other forms of syntax could also be used, e.g. circuit diagrams, maps, flow-charts, chemical formulae, etc.
- We'll define a formal system that includes undefined non-logical symbols that can be used to specify a type of geometry.

To illustrate this, we shall use familiar words, e.g. 'line', 'point', 'intersection', and others, but will pretend that they currently have no meaning except the meaning implicitly defined by their use in a formal theory.
But we will assume the normal meanings of logical operators.

A formal system concerning lines and points

The following are typical axioms used to specify a type of geometry:

UNDEFINED SYMBOLS (with mutually defined types).

Line	[predicate]
Point	[predicate]
Intersection	[function of two lines, returns a point]
Line-joining	[function of two points, returns a line]
Is-on	[binary relation between a point and a line]
Identical	[binary relation]

AXIOMS (A partial set)

If L1 and L2 are Lines and not Identical(L1, L2) then

Intersection(L1, L2) is a Point, call it P_{12}

Is-on(P_{12}, L1)

Is-on(P_{12}, L2)

If Point(P) and Is-on(P, L1) and Is-on(P, L2) then Identical(P, P_{12})

If Point(P1) and Point(P2) and not Identical(P1, P2) then

Line-joining(P1, P2) is a Line, call it L_{12}

Is-on(P1, L_{12})

Is-on(P2, L_{12})

If Line(L) and Is-on(P1, L) and Is-on(P2, L) then Identical(L, L_{12})

This is actually a specification of a class of theories, all of which contain the above axioms but are more specific because they add additional axioms.

We'll give an example later.

The theory in English

Any two distinct lines have a unique point of intersection, and any two distinct points have a unique line containing both.

(Standard projective geometry drops the qualification 'distinct'.)

All of this can be expressed precisely in Predicate Calculus.

Exercise

You may find this theorem easy to prove:

THEOREM

If L1 and L2 are lines, then

Identical(Intersection(L1, L2), Intersection(L2, L1))

In English: the intersection of L1 with L2 is the same point as the intersection of L2 with L1.

A specific projective geometry and its models

Here is an axiomatic specification of a structure with six components, whose axioms are the axioms given previously, plus the following:

SUPPLEMENTARY AXIOMS

P1, P2, P3 are of type Point.
L1, L2, L3 are of type Line.

There are no other lines or points.

Identical(Line-joining(P1, P2), L1)
Identical(Line-joining(P2, P3), L2)
Identical(Line-joining(P3, P1), L3)

Identical(Intersection(L3, L1), P1)
Identical(Intersection(L1, L2), P2)
Identical(Intersection(L2, L3), P3)

It should not be too difficult to produce a drawing containing three blobs and three lines satisfying those axioms, according to the 'obvious' interpretation of the undefined symbols in the axioms: That drawing will be a model for the extended axiom set.

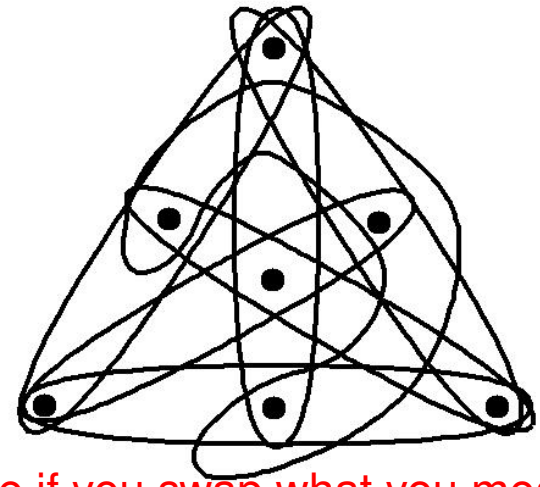
It is worth noting that you can find another model (called the dual model) by treating what we call lines in the drawing as being of type Point and what we call blobs as being of type Line and providing novel interpretations for Intersection, Line-joining, Is-on

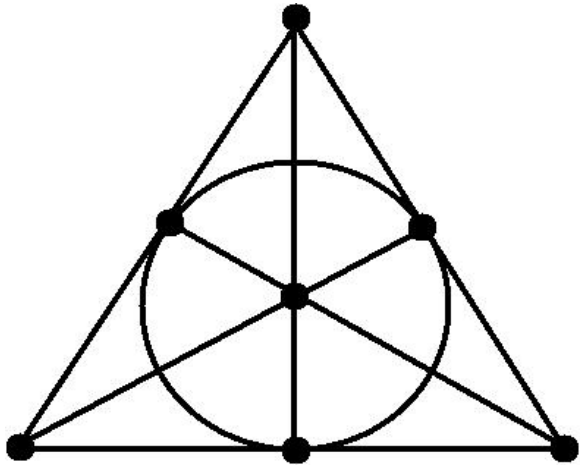
Try it!

Two slightly more complex models

Here are two slightly more complex models of the original axioms, each of which contains seven things we can call points and seven things we can call lines, with obvious interpretations of Intersection and Line-joining, etc., that make all the axioms come out true: Check them out!

Consider carefully what you mean by ‘Point’, by ‘Line’, by ‘Intersection’, by ‘Is-on’, etc. in each case.





In each case if you swap what you mean by ‘Point’ and by ‘Line’, and re-interpret ‘Intersection’, ‘Line-joining’ and ‘Is-on’ you can still make all the axioms come out true.

This is known as the duality of lines and points in projective geometry: anything that is a model for the axioms can be turned into another model by swapping the two sets taken as interpretations for ‘Line’ and ‘Point’, and changing the interpretations of the functions and relations appropriately.

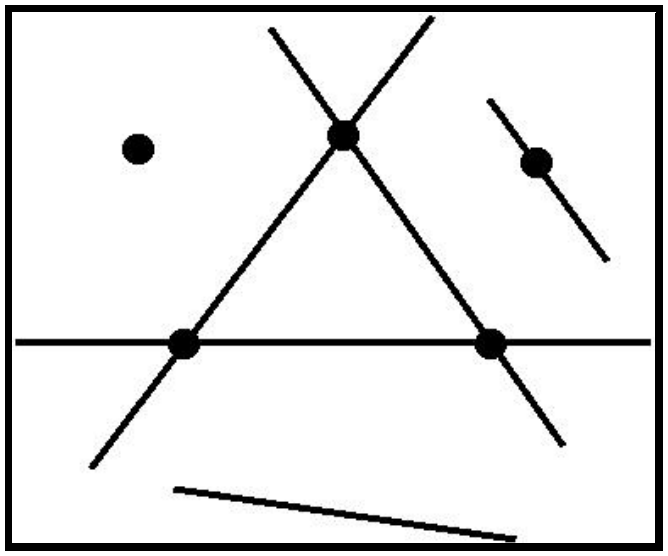
NB:

IF WE ADDED ANOTHER AXIOM, E.G. STATING THAT EVERY POINT IS-ON FOUR DISTINCT LINES NEITHER OF THESE WOULD BE A MODEL:

ADDING AXIOMS CAN REDUCE THE SET OF MODELS.

Some non-models of the axioms

On the right is a picture which if interpreted in the obvious way does not constitute a model of the axioms presented previously.



For example, it is not true that for every pair of points there is a line joining them, and it is not true that for every pair of lines there is a point that is on both of them.

One point is on no lines, and one line has no points on it.

There are many more non-models of the Axioms.

For example, if you have a set of bowls of fruit on a table, you could try mapping 'Line' to the set of bowls, and 'Point' to the set of pieces of fruit, and 'Is-on' to being inside.

But the axioms will not be satisfied, since it is not true that any two "lines" contain a common "point" and it is not true that every pair of "points" determines a "line" they are both "on".

Consistency

If a set of axioms is inconsistent, there will be no models.

If the set is consistent, but the conjunction of axioms does not form a tautology (like (P OR NOT-P AND (Q OR NOT-Q)) then there will be some models and some non-models.

It is possible to invent hypothetical models, e.g. a society in which people are points and social clubs are lines, and any two social clubs have a unique common member, and any two people are members of a unique common social club, etc.

Whether a particular social system is or is not a model of the axioms then becomes an empirical question.

Adding extra axioms, e.g. about properties of points and lines and how they are related, might rule out sociological models

E.g. requiring there to be infinitely many points on each line, etc.

Additional axioms for geometry and motion

NOTE:

The axioms presented above do not fully define projective geometry as studied in mathematics.

The full specification of projective geometry requires additional axioms, e.g. axioms specifying the minimum number of points on a line, and the minimum number of lines through each point.

Further axioms can be added, for instance constraining the points on a line to be totally ordered.

Additional axioms referring to distances and angles are needed to produce a set of axioms for Euclidean geometry.

THE AXIOMS CONSIDERED SO FAR DO NOT MENTION TIME.

- We could add undefined symbols, such as Time-instant, or Time-interval, Particle and At and add axioms allowing particles to be at different points at different times and to be on different lines in different time-intervals.
- Then cars or people moving on roads or trains on railway tracks could be models of the axioms.
- If we added undefined symbols expressing Euclidean notions of distance and direction, and added axioms expressing constraints on how locations, distances, rates of change of distance, etc. can vary, then we could use the axioms to make predictions about particular models.

THAT WOULD REQUIRE US TO ADD SOME LINKS BETWEEN THE THEORY AND METHODS OF MEASUREMENT AND OBSERVATION.

(The theory would need to be 'tethered'.)

Model-theoretic semantics

The idea of a set of axioms having models goes back a long way in mathematics:

Descartes showed that Euclid's axioms for geometry have an arithmetical/algebraic model.

Later it was shown that axioms for the set of positive and negative integers could be satisfied by a model in which integers are represented by pairs of natural numbers (from the set 0, 1, 2, 3,....).

It was also shown that axioms for rational numbers could be satisfied by a model in which every rational number (e.g. 5/8, -16/19) is represented by a pair of integers.

In the 20th century this notion of model was made mathematically precise mainly by the work of Tarski though others contributed including Hilbert, Gödel, Montague and others.

Meaning postulates/Bridging rules

Carnap and others contributed the notion of adding a set of "meaning postulates" or "bridging rules" to a set of axioms: the additions did not define any of the undefined symbols but specified extensions to the axioms which linked them to previously understood notions, e.g. to methods of measurement or experimentation.

This made it possible to draw conclusions like:

- If experiment E occurs then some statements S1, S2, ... expressed using the undefined terms will be true
- If statements S1, S2, ... using undefined terms are true, then some measurement procedure will produce result M.

This enabled theories using undefined terms to be used in making predictions, forming explanations, and deriving tests for the theory.

For example, a physical theory referring to the temperature of objects was tethered by bridging rules specifying ways of measuring temperature.

However those rules do not **define** the concept "temperature" because later physicists can discover more reliable ways of measuring it (more reliable ways of tethering the theory).

The bridging rules do not remove all the ambiguity in the semantics of the set of axioms, but do reduce the ambiguity: the theory does not float quite so free — it is tethered.

Tethering is a way to ensure that the theory's models include a particular bit of the real world and exclude other possible worlds or other parts of this world.

Structure-determined semantics

- The previous slides show that a set of axioms, simply in virtue of its logical structure, if it is internally consistent, allows some structures in the universe to be models and rules out other structures as non-models.
- To the extent that there is a set of models the set of axioms determines a set of possible meanings for all the undefined terms in the axioms.
- The multiplicity of models makes the "theoretical terms" ambiguous, but the ambiguity can be continually reduced by adding more and more axioms, since adding an axiom that is independent of the others (not derivable from them) reduces the set of possible models, by eliminating models that do not make that axiom true.
- Another way of reducing the set of models is by linking one or more of the undefined terms to some concept that already has a meaning (e.g. when we linked 'Line' to our familiar concept of line) or by using Carnapian meaning postulates or bridging rules.
- A major thesis of 20th Century philosophy of science associated with philosophers like Hempel, Carnap, Tarski, and Popper, and independently arrived at by Einstein, is:

  Theoretical terms of science (e.g. "gene", "electron", "valence", "atomic-weight", and many more) are partly defined by their role in theories, and partly defined by "tethering" the theories to particular modes of observation and experiment – not by "grounding" all of the symbols in experience: Most of the meaning comes from the theories.

Processes can also be models

So far, most of the models of sets of axioms have been static structures, but there is no reason why a process, including events and causal interactions, should not be a model of a theory.

(E.g. if the theory includes spaces and times.)

In a process, things can have different properties and relations at different times.

A historical narrative is a formal system, which, if true has some sequence or partially ordered collection of actual events as a model (and perhaps other sequences like that one, in different places at different times), whereas stories that are about invented characters and events are capable of having similar models but may not actually have one in the world as it is and has been.

Typically stories are told on the assumption, usually not made explicit, that characters are constrained by some laws of human behaviour that are not made explicit and that other things are constrained by laws of physics and chemistry, though science-fiction stories often alter the constraints.

Newtonian Mechanics

In Newtonian mechanics there are notions of entities having mass, being acted on by forces, and having locations, velocities and accelerations, where those features can change, but the changes are constrained by "laws of motion".

The theory of Newtonian mechanics has very many types of models of varying complexity.

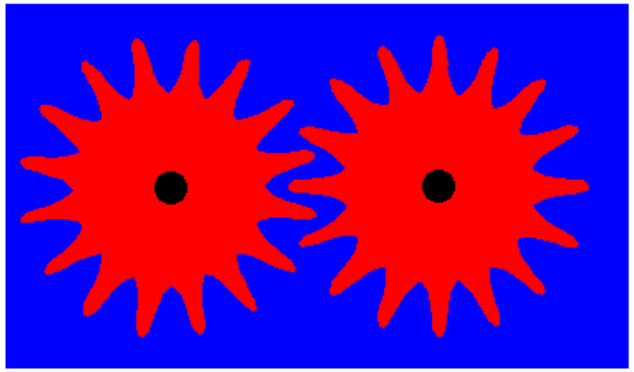
For example, one type of model is a single particle, with no forces acting on it, moving in a straight line at constant speed through an otherwise empty universe.

Another type of model of Newtonian mechanics is a collection of particles one of which has a mass that is much greater than all the others, and they all have motions constrained by mutual gravitational attraction consistent with Newton's laws: our solar system is approximately a model of this type.

Using a process theory to predict

Adding new constraints to a theory restricts the class of processes that model it, so that if a particular observed process P is a model of a theory T, then T can be used for predicting what will happen in P (thereby testing the theory by using it).

If you have a theory about kinds of stuff out of which objects can be made that are rigid and impenetrable, then you can combine that with a theory about shapes in a Euclidean geometry and possible motions in which parts of shapes move while others are fixed, like a wheel pivoted at its centre.
That theory of moving, rigid, impenetrable shapes has very many models, one of which could be a pair of meshed gear wheels each centrally pivoted.

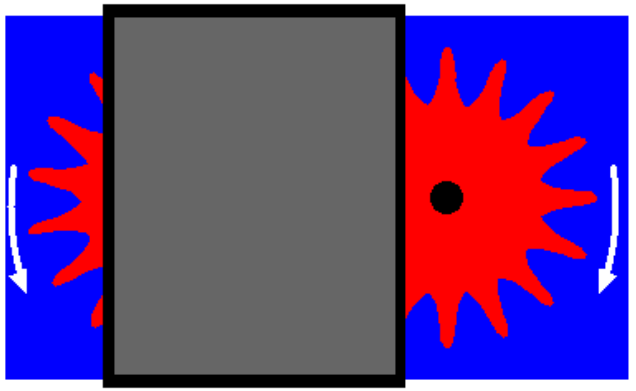


In this case, using the constraints of the theory you can work out that if the left wheel rotates clockwise the right wheel must rotate anti-clockwise – and *vice versa*. (How?)

The theoretical concepts, like Rigid and Impenetrable used in the theory are NOT definable in their full generality in terms of what you can sense (e.g. an invisible object with a shape could be rigid), but their roles in the theory, and their usability in making predictions give them meaning, for a user of the theory able to manipulate forms of representation corresponding to possible models.

Using a theory to explain

The theory can also be used to explain observed processes.



Suppose part of the configuration is covered, and it is observed that when the teeth visible on the left are moved down, the teeth on the right also move down, and reverse their motion when the teeth on the left are moved up.

Building a hypothetical model of the theory, including portions not observed, can allow the observed phenomena to be explained by being predicted using the hypothesised explanatory facts and geometrical reasoning.

Children seem to be able to develop and use such explanatory and predictive theories.

(They use two kinds of causation: Kantian and Humean. See
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#wonac`)

Note

This and the previous diagram are intended to draw attention to the fact that humans and possibly some other animals are capable of using non-verbal, non-logical, non-Fregean forms of representation for reasoning with, as claimed in Sloman 1971).

Elaborations of Model-based Semantics

Note for philosophers and logicians.

Previous slides can be summarised as proposing that semantics can be "model-based and structure-determined".

In his "Two dogmas of empiricism", W.V.O Quine famously produced a description of a scientific theory as a sort of cloud whose centre has little contact with reality, and which can only be tested empirically at the fringes. `http://www.ditext.com/quine/quine.html` (See Section VI.)

This is a suggestive exaggeration.

F.P.Ramsey pointed out that there is a way of removing undefined symbols by treating them as variables bound by existential quantifiers (using a higher order logic).

In this way a theory expressed in predicate calculus can be transformed into a single complex, existentially quantified, sentence: a Ramsey Sentence.

Discussion of Ramsey's idea is beyond the scope of this presentation.

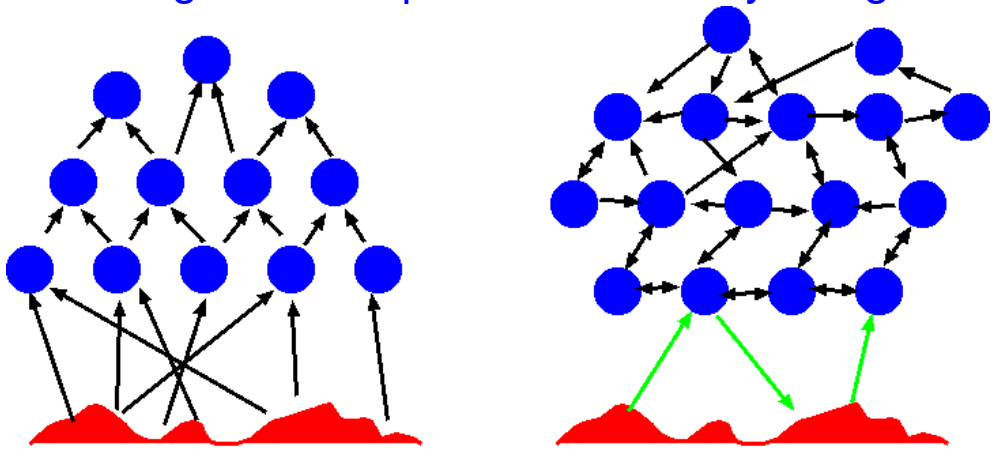
For more on Ramsey, Carnap and others see the lecture by Michael Friedman 'Carnap on Theoretical Terms: Structuralism without Metaphysics', included as part of this workshop:

`http://www.phil-fak.uni-duesseldorf.de/en/fff/workshops/tfeu/` Theoretical Frameworks and Empirical Underdetermination Workshop
University of Düsseldorf, April 2008.

If I am right in claiming that the introduction of theoretical terms by scientists and the substantive ontology extension by pre-verbal children and some other animals achieved as part of exploring how the world works has much in common with how science works, then if the children and other animals are not using human languages or anything like predicate calculus, that implies that pre-verbal children and other animals have other formalisms (e.g. formalisms that may be more like pictures than like sentences?) that are capable of expressing meanings that are partly implicitly defined by the structures used. Pictures of meshed gears may be examples. See the slide below: Beyond logical (Fregean) formalisms

Symbol grounding vs Symbol tethering pictured

The distinction we are making can be represented crudely using this diagram.



On the left every symbol gets its meaning from below (i.e. built up from sensory, or sensory-motor, information). On the right much of the meaning of a symbol comes from its role in a rich theory, expressed in a formalism that allows conclusions to be drawn in a formal way. Tethering merely adds further constraints that help to pin the meaning down.

Note that as a theory develops over time, new modes of experimentation and measurement may be developed. But the same concepts may continue to be used, even though they are linked to observation and experiment in new ways, usually making possible far more precise and reliable measurements and testing procedures than previously.

The idea that the concept (e.g. 'charge of an electron') is preserved across such changes would make no sense if the concept were defined by its links to observation and measurement, since they keep changing.

See also: L.J. Cohen, 1962, *The diversity of meaning,* Methuen & Co Ltd,

Beyond logical (Fregean) formalisms

Formal work on model-based semantics normally considers theories expressed using logical formalisms, or more generally Fregean formalisms (i.e. those using syntax based on application of functions to arguments).

A distinction was made between Fregean and analogical representations in 1971:
`http://www.cs.bham.ac.uk/research/projects/cogaff/04.html#analogical`
arguing that analogical representations can also support valid forms of reasoning, despite using diagrams, pictures and models instead of logical formalisms.

For example, diagrams are often used to work out the effects of lenses, mirrors and prisms on the propagation of light, and chemical formulae can be regarded as diagrams of a sort, used in predicting results of or constraints on chemical reactions.

So the use of theories in constraining meanings of undefined terms is not limited to the use of symbols in logical and algebraic expressions.

However the theory of how such non-logical formalisms can constrain models is not so well developed, despite the wide-spread use of circuit diagrams, flow-charts, maps and many other non-logical notations in science and engineering.

At this stage it is not at all clear what sorts of formalisms are available to infant humans and other animals that are generating new theories about the contents of the environment, and extending their ontology, as they play and explore and try to cope with surprises, by bootstrapping their knowledge about the world: but they probably don't use predicate calculus, e.g. in coming to understand gear wheels.

Theories about non-physical processes

When the environment contains information-processing systems, e.g. people, theories about information-processing mechanisms are useful.

Typically, what is important in such theories is what information is available, how it is acquired, transformed, stored, interpreted, used, etc., as explained in:

`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/whats-information.html`

It is often useful to describe those processes independently of the physical mechanisms that make them happen, i.e. by describing a virtual machine.

We do this when we specify computer software, or describe what's going on in running software systems (e.g. operating systems, compilers, spelling checkers, email systems, databases, games, etc.)

We also do this when we describe ourselves, or other people as having beliefs, desires, preferences, skills, etc., and when consider processes such as decision making, plan formation, becoming puzzled, noticing things, etc.

Humans have an innate self-extending mechanism for generating theories about virtual machines in other agents (evolution 'invented' the idea of virtual machines before we did).

At first the theories are very impoverished, but various kinds of observation of self and others, and social interactions seem to drive a process whereby theories of mind are extended with new theoretical concepts and new constraints about how things work (sometimes false ones).

Concepts of beliefs, percepts, intentions, attitudes, skills, etc. cannot be "grounded" in sensory or motor signals: they refer to virtual machine states and processes.
Mental concepts, like physical, chemical, biological concepts, get most of their semantics from the theories in which they are used, which develop slowly over several years.

Reducing semantic indeterminacy

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

Perhaps Kant's "things in themselves"?

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.
Both adding more axioms and adding new attachment points can reduce the semantic indeterminacy of the "undefined" symbols.

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 instead of assuming infant learners are ham-strung by the need for all their internal symbols to be "grounded" by being defined in terms of sensory and motor signals.

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.)

Making these ideas work

So far, the presentation has been purely theoretical, showing how in an abstract sense axioms can constrain possible models, by excluding some portions of the universe, or some mathematical structures that do not satisfy the axioms (under specific interpretations of the undefined symbols) and including others.

A major task remaining is to show how a developing animal or robot can make use of that abstract possibility in order to simultaneously extend its ontology (its set of concepts describing things that can exist) and its theories that make use of that ontology, in perceiving, describing, explaining, predicting, planning, and producing states of affairs in the world.

This will require significant extensions of robot designs beyond the current technology, which typically assumes all concepts that are used in the contents of percepts, thoughts, goals, etc. are tied to sensory mechanisms, or sensory-motor control subsystems.

This presentation says nothing about how to proceed beyond such designs. I have some ideas but will expand on them in a separate document.

A robot should not be constrained to do all of its theorising using logic: we need to explore a host of forms of representations and mechanisms for constructing, transforming, and deriving representations, with different properties.

An attempt to do this that has not been widely noticed is presented in Arnold Trehub's book *The Cognitive Brain*, MIT Press, 1991, online here: `http://www.people.umass.edu/trehub/`

An exception: Map-making robots

It is worth noting that an exception in the current state of the art is the representation of rooms, walls, corridors and doors in SLAM (simultaneous localisation and mapping) systems.

- The robots involved in those systems typically represent the layout using a map-like structure (with topological and possibly also metrical properties) whose ontology is exo-somatic (i.e. refers to things that exist independently of any sensor and motor signals or other internal states and processes of the agent) and in particular is not **defined** in terms of somatic sensory-motor relationships.
- However the map-like structure can be linked by suitable perceptual, learning, planning, and plan-execution mechanisms to sensory and motor signals – a form of tethering.
- Such tethering plays a role in growing or modifying the maps and also enables the maps to be used for predicting consequences of movements and consequences of change of gaze direction, and for planning routes and controlling actions when executing plans.
- The sort of representation and ontology typically used in such SLAM systems does not satisfy the strict requirement for symbol-grounding, for a representation is constructed and used that is neutral as to the particular sensory and motor signals of the user.
- For example the same map-structure could in principle be shared by two mobile robots with different sensory and motor sub-systems, e.g. one with TV cameras and wheels, and the other with laser sensors and legs.

What do we need to explain?

Besides demonstrating ways in which "ungrounded", but suitably "tethered" forms of representation can be used in intelligent machines we also need to explain how they could come to be used in biological organisms.

This has at least three distinct aspects

- Explaining the functional need for mechanisms using such information structures,

  E.g. identifying features of a niche that determine requirements that such mechanisms can meet:
  - Providing innate conceptual apparatus for perceiving and learning (Kant)
  - Making predictions at a high level of abstraction
  - Formulating goals and plans
  - Recording re-usable information about the environment, including spatially or spatiotemporally specific information and useful generalisations

- Explaining how such mechanisms and forms of representation could evolve in species that do not yet have them.

  E.g. studying trajectories in niche space and in design space.

- Explaining how such mechanisms and forms of representation could arise during learning and development in individuals.

  Including studying both the features of the environment that provide a need and opportunities for such changes, and the mechanisms in individuals that make such changes possible.

Giovanni Pezzulo & C. Castelfranchi, "The Symbol Detachment Problem", in *Cognitive Processing* 8, 2, June, 2007 pp 115-131. http://www.springerlink.com/content/t0346w768380j638/

Clayton T Morrison, Tim Oates & Gary King, "Grounding the Unobservable in the Observable: The Role and Representation of Hidden State in Concept Formation and Refinement", *AAAI Spring Symposium* 2001, http://eksl.isi.edu/files/papers/clayton_2001_1141169568.pdf.

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.

Meanings are primarily specified by formalisms used to express them and mechanisms that manipulate the formalisms: links to reality added through transducers merely help to refine partially indeterminate meanings.

Note that much creative story-telling referring to non-existent things would be impossible if we were restricted to using only concepts abstracted from sensed things in the environment.

All this is a refinement of the theory in two old papers available at

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

- IJCAI-85 “What enables a machine to understand?”
- ECAI-86 “Reference without causal links”.

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 number-concepts are multi-faceted concepts?

See chapter 8 of The Computer Revolution in Philosophy.
`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`

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 be rejected and replaced 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, and that process seems to be able to go on indefinitely.

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 an experience-grounded 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)

Meaning postulates

As reported above, 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.

Some wild speculations about colour concepts

The previous discussion suggests that grasping colour concepts may in part be a matter of plugging certain parameters (perhaps procedural parameters) into a mostly innately determined complex mechanism encoding an implicit theory about the nature of object surfaces in our 3-D environment — 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 much 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 sometimes be a small role!

Generalising the notion of “language” or “formalism”

The model-theoretic ideas used to introduce claim that a set of meanings, or an ontology, can be at least partially determined by the structure of a formal theory, uses research by logicians and philosophers of science that assumed that all theories were expressed using something like predicate calculus: an assumption that should be abandoned in the context of non-human organisms and pre-linguistic children.

The idea of a “generalised language” (GL) that evolved before human communicative language and develops earlier in children is introduced in this presentation:

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

It is argued there that two key properties of human languages, namely structural variability and compositional semantics need to be generalised and extended to other forms of representation, including spatial representation, such as diagrams, maps, flow-charts.

The results of model theoretic semantics will not immediately transfer to these languages, since their syntactic forms are very different and the modes of composition leave far more scope for context to determine interpretation.

So an open research project is exploring whether and how non-logical languages can support development of new ontologies and new explanatory and predictive theories using those ontologies: This will be relevant to learning and developing in humans and some other animals, and in future robots.

How is ontology-extension controlled?

I have claimed that substantively new concepts can be acquired by a scientist or a child by developing new theories that make use of new concepts expressed by undefined symbols, but controlling the search for good new concepts and axioms is a major problem.

Adding a new hypothesis or axiom to a theory for the purpose of explain already known facts is a process known as “abduction”.

Usually abduction is assumed to use pre-existing concepts, whereas my claim is that there can be ontology-extending abduction too.

> Abduction is often incorrectly presented as a third form of reasoning alongside deduction (working out what must be the case, given some premises) and induction (working out what is likely to be the case, given some premises that do not suffice to support deductive inferences).

Abduction is not a process of drawing some conclusion, but a process of formulating a conjecture which can be combined with other assumptions in a process of reasoning, e.g. to explain or predict something.

Abduction involves selecting candidates from a search space, e.g. the space of possible hypotheses (or geometric models) that could usefully and consistently be added to what is already known.

In general that is a huge space, and if the search is not constrained in some way it can be impossibly difficult.

The roles of abduction in philosophy and science are compared and contrasted here:

`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html`

For more on abduction and ontology extension see

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#mofm-07`

(Some relevant papers on multimodal abduction by Lorenzo Magnani are mentioned at the end of those slides.)

Acknowledgements

An earlier version of this was 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.

More recently answering Jackie Chappell's questions helped to refine the ideas and led me to switch terminology from "symbol attachment" to "symbol tethering", and pointed towards some ideas about possible implementations. (Not soon.)

This 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.

I hope it will not be too long before the ideas can be implemented in a learner that shows how the use of model-based semantics with symbol tethering renders symbol-grounding unnecessary!

Some references

Some of the ideas are in my 1986 paper on “Reference without causal links”:

http://www.cs.bham.ac.uk/research/cogaff/Sloman.ecai86.pdf

and in my 1978 chapter on the aims of science, now available in this free online Book (The Computer Revolution in Philosophy):

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

The Birmingham Cognition and Affect Project and the CoSy project

PAPERS (mostly postscript and PDF):

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

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

Especially the work with Chappell, and criticisms of sensorimotor theories.

(References to other work can be found in papers in those directories)

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/research/projects/cogaff/talks/

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

Comments, criticisms and suggestions for improvements are always welcome.