The Design-Based Approach to the Study of Mind (in humans, other animals, and machines), Including the Study of Behaviour Involving Mental Processes.

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

Installed: 2005
Last updated:
7 Mar 2009; 30 Jan 2010; 28 Feb 2010; 6 Sep 2014; 28 Jan 2016; 31 Dec 2017 (references, format)

This paper is available in HTML and PDF formats:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/design-based-approach.html
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/design-based-approach.pdf

---

CONTENTS

What is the design-based approach ("designer-stance")

When scientists attempt to explain observations of behaviour in humans and other animals, they often use language that evolved for informal discourse among people engaged in every day social interaction, like this:

- What does the infant/child/adult/chimp/crow (etc) perceive/understand/learn/intend (etc)?
- What is he/she/it conscious of?
- What does he/she/it experience/enjoy/desire?
- What is he/she/it attending to?
- Why did he/she/it do X, start Xing, stop Xing, speed up Xing... ?
- Does he/she/it know that ...?
- What did/does he/she/it expect will happen, if...?

Similar comments can be made about the terminology used in many philosophical discussions about minds, cognition, language, and the relationships between evolution and learning.

These forms of description and explanation make use of a collection of theoretical assumptions and strategies similar to what Dennett called "the intentional stance" and Newell called "the knowledge level" (the differences need not concern us now). That approach treats all those whose behaviour is being explained, whether animals, infants, toddlers, and in some cases people with serious psychiatric disorders, as if they were all basically like normal human adults in the way they operate, taking decisions and acting on the basis of what they know, what they perceive, what concepts they have, what goals, preferences and attitudes they have, and how they reason, deliberate and plan. Sometimes we can also invoke ways of being irrational, for example when experiencing strong emotions, though it is not clear that that would be included in Dennett's "Intentional stance".

There is nothing wrong with that if the aim is to inform, entertain, speculate, educate in a general way, generate interest, influence policy making, or tell creative stories. However, a different approach is needed if the aim is to provide *scientific* understanding: the kind of understanding of how humans and other animals work that could enable us to explain what they can and cannot do, understand how they learn and develop, or how their development can go wrong, and if we wish to gain insights into how they evolved, and the relationships between evolution and development.

Going beyond these 'common sense' descriptions of animal behaviours and competences requires us to formulate theories about the mechanisms within the animal that produce the behaviours.

Sometimes scientists (or philosophers) attempting to produce explanations of the phenomena observed, or hypothesised, try to describe what is going on *inside* the person or animal. But the ways they have of doing that derive from concepts used in everyday conversation for describing human mental states and processes, such as *noticing, seeing, expecting, deciding, comparing, choosing, learning, hypothesising, wanting, preferring,* and many more.

One characteristic of the above concepts is that they normally refer to what a whole person is doing, e.g. you or I notice something, not a bit of your or my brain or one of our eyes or ears. But when scientists feel that that is not adequate as an explanation they try to identify either *physical* parts of a person or animal, or parts labelled in terms of their cognitive functions, whose operations are supposed to explain what happens.

What sorts of parts/components should be referred to in adequate explanations is not easy for some researchers to understand.

If we are explaining the behaviour of car or clock, we think the parts that are relevant to explaining what happens are physical components that can be identified separately from other components, and which do specific things they were designed to do. If we wish to explain what happens when a volcano erupts, or chemicals react, or a plant grows we also refer to interacting physical parts, though without assuming they were designed by humans or any other intelligent designer to do to anything.

That strategy of explaining in terms of interacting physical parts is very successful in the physical sciences, and in many branches of engineering (including explaining malfunctions in machinery as well as how things work). So there is a strong temptation to look for physical parts to explain human and animal competences and behaviours, and typically that involves trying to find which bits of brains are relevant, along with which bits of bodies (sensors and effectors).

Physicists and chemists have learnt a lot about the items involved in physical and chemical interactions, and engineers often know a lot about what the parts of the machines they build can do in various circumstances.

However brains are far more complex and obscure in their operation than any of the complex systems studied by physicists and chemists or the machines designed by engineers. It isn't even clear what most of their functioning components are or what their functions are, though large numbers of fragmentary discoveries are being made about which bits seem to be involved in which processes, and about how the parts interact physically and chemically.

So on the one hand we get neuroscientists listing components whose ability to produce processes like perceiving, deciding, learning, hypothesising, planning, wanting, evaluating is largely mysterious, and on the other hand we get behavioural and cognitive scientists, and even AI theorists, listing hypothesised components that are often described using familiar common sense concepts, like *perceiving, deciding, learning, ... evaluating* where the concepts are very loosely defined in a scientific explanatory context (though not in ordinary conversation) and their use in explanations is *circular* because the concepts already presuppose that these systems have capabilities of the sorts we want to explain.

It's as if someone tried to explain how a car engine works by listing and labelling parts without indicating how any of them work or how they interact, e.g. by saying, this is the bit that starts the car, this is the bit that makes the car go faster, this is the bit that makes the car slow down, etc. This leaves the task of explaining how any of those parts do what they are assumed to do.

Towards more satisfactory theories

I think developments in computing, AI, and computational cognitive science, during the last half century have taught us that there is a mode of explanation of how complex systems work that is very different from *both* describing their physical, chemical, electrical, or mechanical parts and operations *and* describing them using common mentalistic language. That new alternative uses what we have learnt about designing complex information processing systems of many kinds, though none come near the specific kinds of sophistication that we wish to explain in humans (young and old) and other more or less intelligent animals.

The particular forms of explanation refer to such things as

- The kinds of information that the organism has.
- How the information is acquired (including which features of the environment make it available and how the sensory and perceptual mechanisms acquire it.
- The various ways in which the information can be manipulated, analysed, recombined, derived, or used in planning and problem solving.
- The forms of representation used to encode that information and the mechanisms that operate on those forms of information.
- The ontologies that constitute the basic information structures the organism uses, from which more complex information is constructable.

- The architectures in which those various capabilities, mechanisms, and information structures are combined.

- The processes by which all the items listed above (kinds of information, forms of representation, mechanisms, architectures, etc.) are initially constructed and processes by which they continue to grow and develop over extended periods in some organisms (e.g. humans).

- The kinds of information, forms of representation, mechanisms and architectures that need to exist at a very early stage to support all those developments.

- The ways in which the nature of the environment constrains the types of information that are available to the organisms and poses problems that the information needs to be used to solve. (For two different species living in the same location, the role of the environment can be very different, because of the effects of their different evolutionary past, producing different bodily forms, different needs, different forms of reproduction, etc., using different sorts of information.)

We need to understand that in talking about a mind (or a major subsystem of a mind) we are talking about a complex system with many concurrently active parts -- that work together more or less harmoniously most of the time but can sometimes come into conflict, whose working need to be understood in terms that can help to bridge the gap between the functions they are known, or assumed, to have, and the underlying mechanisms that make it possible to have such functions and which limit and shape those functions.

These parts are organised in an information-processing architecture that maps onto brain mechanisms in complex, indirect ways that are not well understood.
So, when studying some human (or animal) psychological capability or limitation, we should ask questions like this if we wish to do deep science:

- Which parts of the architecture are involved?
- What are their functions?
- What kinds of information do they acquire and use?
- How do they do this?
- What is the total architecture in which the various parts function?
- How is the information represented?
  (It could be represented differently in different subsystems for different purposes).
- What kinds of manipulations and uses of the information occur?
- What mechanisms make those processes possible?
- How are the internal and external behaviours selected/controlled/modulated/coordinated?
- How many different virtual machine levels are involved and how are they related (e.g. physical, chemical, neural, subsymbolic, symbolic, cognitive,...)?

One of the current fashions straddling philosophy and cognitive science emphasises intelligence (or mentality, or consciousness) as "embodied" and "situated". The people who defend such theories or methodologies have mostly failed to understand at least half the points listed above, as shown by the examples they use (an extreme case being a "passive walker" robot that doesn't need a brain to walk down a slope, but which cannot walk up the slope, nor cope with a brick obstructing its downward path. Compare the "Chewing Test" for intelligence, presented here (since 2014):
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/chewing-test.html

Different requirements and different designs

It is important that in studying those questions we remember that humans (and other animals) evolved to function in a particular type of environment found on this planet. Moreover there are difficult solutions to different subsets of the problems of surviving in such an environment. A deep understanding of one design would include being able to compare and contrast it with other designs and with different sets of requirements and constraints imposed by the environment (including other animals).

Much writing in philosophy and psychology falls into the trap of assuming there is only one way of perceiving, wanting, believing, deciding, and being conscious -- the human way. In trying to describe the human way they often make the mistake of forgetting that we are products of evolution and there were many different distinct designs in our precursors. If we can find theoretical or empirical evidence for some *alternative possible* requirements and designs, we'll be in a far better position to understand the detailed structures and the costs and benefits of the *actual designs* found in humans.

When presenting theories about cognitive functions and the mechanisms that explain them it is very important to try to be clear about the precise collection of requirements which the proposed mechanisms are supposed to meet. In particular, we need to distinguish at least the following, though far more fine-grained distinctions between requirements will be needed.

Kinds of requirements

- Producing behaviour in real time that is suited to the precise configurations of things in the environment that define the goals, and the positive and negative affordances.
- Using sensed/perceived information immediately in control of actions or decisions (**online intelligence**) and storing sensed/perceived information in a format, or collection of formats, that can be useful later in a variety of contexts, including context not anticipated when the information was acquired (**offline intelligence**).
- Thinking, reasoning, explaining or making plans concerning actions that are not currently being performed but which could be performed in the future, or were performed in the past (by the person or animal concerned).
- Perceiving thinking, reasoning about, perceived processes occurring in the environment not caused by the individual, but which may or may not affect the individual (proto-affordances), or may be relevant to the goals or actions of another individual (vicarious affordances).
- Perceiving thinking, reasoning about, the percepts, thoughts, desires, plans and actions of another intelligent agent, as opposed to something like wind, water or gravity that can cause things to happen without using any cognitive mechanisms. Being able to perceive, think, reason, or deliberate about other individuals with similar powers requires **meta-semantic (and in some cases meta-meta-semantic) competences**, which not all organisms seem to have.
- Being able to use meta-semantic competences directed at one's own thinking, reasoning, perceiving, etc., for instance finding a flaw in one's planning strategy and repairing it. This requires self-directed meta-semantic competence.

There are also deep problems about how many of these different competences can operate concurrently, so that there is no

```
sense->think->decide->act
```

cycle as is sometimes assumed, but a number of different subsystems running and interacting (including self-monitoring subsystems), as illustrated by the CogAff architecture schema and its H-Cogaff special case
http://www.cs.bham.ac.uk/research/projects/cogaff/#schema

Kinds of Observations Needed
A task on which more thought is required is how the research goals listed here can influence choice of experiments and the observations required.

One implication that is not obvious is that insofar as different individuals have different combinations of knowledge, concepts, forms of representation and possibly also architectures (e.g. if they are at different stages of development), important information may be lost by focusing on averages across collections of experimental subjects, as opposed to adopting a "clinical" approach and trying to describe in detail what exactly different individuals do and what that implies regarding differences in how they do things. This can also be important in making studies of development more fine-grained than is common when averages across populations in a species or in an age group are used (often without even paying attention to the variance!)

The switch from common-sense descriptions to descriptions that are useful in designing working computational models is not always an improvement, since sometimes the ontology available to the designers is too restrictive. A simple example concerns the type of programming language used. Some programming languages, especially the earliest ones, were aimed almost exclusively at specifying numerical computations, whereas from the beginnings of AI it was clear that computers would have to manipulate non-numerical information structures, including sentences, logical expressions, grammars, parse trees, plans, equations, and various structures built from those. AI languages like Lisp, Pop-11, Prolog, Scheme and others were designed accordingly. However there are many AI/Robotics researchers who are unfamiliar with such languages and whose programming skills are most geared to numerical computations. As a result such designers develop systems in which all information about sensory input, about structures and processes in the environment, about motivation, and about control of actions is expressed in numerical form, including, for example, the use of a global cartesian coordinate system to represent spatial locations, orientations, distances, sizes, and relationships -- a practice criticised in this presentation
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#babystuff
Ontologies for baby animals and robots: From "baby stuff" to the world of adult science: Developmental AI from a Kantian viewpoint.

A consequence of adopting this approach

If we think of features of humans and other animals such as consciousness, intelligence, attention, memory, emotions, autonomy in this ’design-based’ way (adopting what John McCarthy also now calls ’the designer stance’) the sorts of questions we can ask and the sorts of theories we can consider are expanded in an important way.

A design for a working system (microbe, ant, chimpanzee, human, robot) will specify a complex virtual machine with many coexisting, interacting information-processing components, as explained in this PDF presentation:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk51
Why robot designers need to be philosophers -- and vice versa.
And this introduction to Virtual Machine Functionalism:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html (also PDF)

Since there are many components, it is possible to consider different designs for working system that use different combinations of such components, and different versions of the components. This sort of variation in designs is evident in the products of biological evolution.

A corollary is that where we are naturally inclined to think of a *binary* division such as a division between animals that do and do not have some feature X (consciousness, creativity, autonomy, emotions, planning capability, free-will, etc.) the design based approach replaces the binary division by something more like a *taxonomy* or a *grammar*, allowing for a (possibly large) collection of cases where there are typically many discontinuities

> **Example:**
> Many people (including me once) assume that there is a binary division between *reactive* and *deliberative* control mechanisms.
>
> After hearing several presentations and reading several documents making use of these labels in confusingly different ways, I eventually realised that people were interpreting the division in different ways because the space of possible designs had a kind of complexity that had not been studied properly and people were basing the distinction on different 'cracks' in the space.
>
> For example, some people were using 'deliberative' to refer to any system that could, in some sense, evaluate alternative action possibilities and select one, whereas others used the label to refer to more complex systems that can plan more than one step ahead when taking decisions.
>
> When I looked closely, I found that there were several more important sub-divisions between different sorts of deliberative competence, and attempted to document them in this discussion note on 'fully deliberative' systems:
> http://www.cs.bham.ac.uk/research/projects/cogaff/misc/fully-deliberative.html
>
> I don't claim that the analysis of that document is complete: there may be more sub-cases to distinguish.

When considering any X and asking which animals have X, how X evolved, what X's costs are, what Xs benefits are, which neural or other mechanisms are involved in X, etc. a good heuristic is to ask

- How many varieties of X are there?
- What sorts of distinct components, that might have evolved separately, are involved in different varieties of X?
- How do different combinations of components affect functionality of X?
- How might those components have evolved?

See the Turing-inspired Meta-Morphogenesis project: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html (also PDF)

And that generally has the result of replacing a binary divide between things that do and do not have X with a collection of varieties of X, and a collection of intermediate cases between not having a particular sort of X and having it.

This idea was used in an analysis of the notion of 'free will' originally posted to a 'usenet' news group, now available here: http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#8

(Added 30 Jan 2010)Relevance to ontology/ontologies

The previous section pointed out a consequence of this study of varieties of architecture that might have been designed or might have evolved in response to explicit or implicit requirements such as the pressures of an ecological niche, namely that we usually need a richer ontology of types of design than can be expressed using binary distinctions normally assumed.

Another consequence of the approach is that the process of designing a complex working architecture, testing it and finding problems that need to be addressed by improvements in the design, often teaches us that there is a much richer variety of possible internal states than we might have considered possible in advance.

Systems with complex virtual machines that include concurrently active interacting subsystems, including some subsystems that monitor and control others, can have a richer variety of internal states and processes than can be defined in terms of varieties of external behaviour, or even relations between inputs and outputs. For example, a system can run internally and have no connections with output signals most of the time, even though it occasionally is linked to inputs and outputs.

A simple example could be a complex virtual machine that is capable of playing many different games, and at any time practices some of those games internally by playing itself, e.g. at chess, or draughts (checkers), as a result of which its competence in those games increases, though there is no external sign of those changes unless it engages in a game with an external player, which may never actually happen.

Its input and output channels may have capacity limits that limit the total number of games that the system actually plays in its lifetime, and that limit may be significantly lower than the number of different games it has the competence to play.

By studying the variety of internal states that the architectural design (the information-processing architecture) of some organism makes possible we may find that to understand and explain how the organism works we need a much richer ontology of states and processes than would be suggested merely by watching its behaviours and trying to classify them.

This is particularly rue of humans: there is no reason to suppose that the ontology expressed in our ordinary language concepts for talking about mental processes, or even the extensions to that ontology developed by psychologists and psychiatrists as a result of interacting with

and experimenting on humans is rich enough to account for all the important phenomena of human life: instead we need a much richer ontology of states and processes derived from a good theory of how the system works. This is similar to the way our understanding of the variety of types of material substance had to be substantially revised when we discovered the underlying architecture of matter, as composed of atoms of various sorts that can combine to form molecules of various sorts that can be arranged in configurations of various sorts -- none of which was dreamt of prior to the development of modern physics and chemistry.

A relevant discussion paper

The issues raised here are pursued further in different ways in different online papers produced as part of the Cognition and Affect project and the CoSy Robot project, referenced below.

A particularly relevant methodological discussion paper is

> Two Notions Contrasted: 'Logical Geography' and 'Logical Topography' Variations on a theme by Gilbert Ryle: The logical topography of 'Logical Geography'.
> http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html

That paper discusses relationships between philosophy and science in the context of an attempt to clarify Ryle's notion of 'Logical Geography', showing that there is a deeper type of investigation, which I called the study of 'Logical Topography', which identifies aspects of some portion of reality that allow various possible kinds of concepts to be developed, in contrast with the study of the concepts that are *actually* in use, Ryle's 'Logical Geography'.

The difference emerges in two ways: The study of logical geography assumes (a) that there is one collection of concepts whose relationships can be charted and (b) that this will answer philosophical questions definitively. The study of logical topography reveals (a) that the relevant aspect of reality can be divided up in different ways, leading to different logical geographies, and (b) that that reality may itself may have unnoticed complexity of structure, which, when explored in depth, shows possibilities that were not exposed by the original philosophical investigations.

On the basis of those ideas, the paper identifies a kind of philosophical theory building that has much in common with scientific theory-building (including the ability to introduce extensions to our ontology), and which uses abduction.

See also

- The Cognition and Affect Project and our papers in the CoSy project.
- What's Information? (HTML)
  A discussion paper on the nature of information, and why the concept of information is as important for science (and engineering) as the concepts of matter and energy.

- Why robot designers need philosophers and vice versa (PDF)
  Short tutorial presentation at University of Bielefeld (Oct 2007).
  (Among other things explains why the notion of 'virtual machine' is essential to the study of mind.)

- The Mythical Turing test

- Other online presentations (mostly PDF)

Originally installed here in 2005
Some of the ideas were in this paper

> Prolegomena to a Theory of Communication and Affect
> In Ortony, A., Slack, J., and Stock, O. (Eds.) *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues.*
> Heidelberg, Germany: Springer, 1992, pp 229-260.

It distinguished various approaches to the study of mind: the design-based approach, the phenomena-based approach, and the semantics-based approach. I could have added more, e.g. the mechanism-based approach that starts from assumed 'known' mechanisms (e.g. neural nets) and tries to show how they can account for the observed, or introspected, phenomena. Daniel Dennett's "intentional stance", also referred to as "The knowledge level" by Herbert Simon and Allen Newell, could be mentioned here: it assumes that the individuals being studied are rational. I regard that assumption as both unjustified and unnecessary for studying and modelling biological information-processing systems (e.g. humans, ants, microbes).

The design based approach is closely related to the study of logical topography of sets of concepts, as opposed to the logical geography.
See http://www.cs.bham.ac.uk/research/projects/cogaff/misc/logical-geography.html

Using a "design-based" approach is partly similar to adopting what Dennett has called the "design stance", in at least some interpretations of what he is saying, namely an interpretation in which one considers *How something works* and h *How other things like it could be made.*

There is a different interpretation of Dennett's design stance, summarised here

> http://en.wikipedia.org/wiki/Intentional_stance
> When we predict that a bird will fly when it flaps its wings, on the basis that wings are made for flying, we are taking the design stance. Likewise, we can understand the bimetallic strip as a particular type of thermometer, not concerning ourselves with the details of how this type of thermometer happens to work.
>
> On that interpretation one only assumes that something has been designed (possibly by evolution) to perform a certain sort of function and uses that assumption to predict what it will do. That does not use what I have called "the design-based approach", which involves trying to understand how it works. I suspect Dennett switched between the two interpretations.

The main difference is that what I have called the "design-based approach" emphasises the need for *comparative* investigations in different parts of both

- the space of possible designs and
- the space of possible sets of requirements.

This comparative approach includes comparing different requirements and designs found in biological organisms and their niches. Another possible difference arises from an ambiguity in some of Dennett's presentations. He can be read as suggesting that the "design-stance" is concerned with the physical design of something, e.g. concerned with what the electronics and mechanics do and how they do it, in the case of artifacts, and concerned with what the physiology does and how it does it, in the case of organisms. In contrast my emphasis is mostly concerned with the design of

information processing systems, of which the majority that are of interest are virtual machines with information-processing functions, as opposed to physical components with functions. (Sometimes Dennett seems to include this, though he sometimes seems to me to imply that talk of virtual machinery requires the intentional stance, rather than the design stance.)

I believe John McCarthy's use of the term "the designer stance" is closely related to what I have called "the design-based approach".
See footnote 5 in his paper "The Well-Designed Child", originally written in 1996 then published, with minor changes, in the *Artificial Intelligence journal in 2008*
http://www-formal.stanford.edu/jmc/child.html
Alongside my paper "The Well-Designed Young Mathematician"
http://www.cs.bham.ac.uk/research/projects/cogaff/08.html#807
Compare my Kantian Philosophy of Mathematics and Young Robots.

Maintained by Aaron Sloman
School of Computer Science
The University of Birmingham