Proceedings of the International Symposium on AI-Inspired Biology, Jackie Chappell, Susannah Thorpe, Nick Hawes and Aaron Sloman (Eds.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

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[1]
http://www.cs.bham.ac.uk/~axs

Abstract. There is much work in AI that is inspired by natural intelligence, whether in humans, other animals or evolutionary processes. In most of that work the main aim is to solve some practical problem, whether the design of useful robots, planning/scheduling systems, natural language interfaces, medical diagnosis systems or others. Since the beginning of AI there has also been an interest in the scientific study of intelligence, including general principles relevant to the design of machines with various sorts of intelligence, whether biologically inspired or not. The first explicit champion of that approach to AI was John McCarthy, though many others have contributed, explicitly or implicitly, including Alan Turing, Herbert Simon, Marvin Minsky, Ada Lovelace a century earlier, and others. A third kind of interest in AI, which is at least as, old, and arguably older, is concerned with attempting to search for explanations of how biological systems work, including humans, where the explanations are sufficiently deep and detailed to be capable of inspiring working designs. That design-based attempt to understand natural intelligence, in part by analysing requirements for replicating it, is partly like and partly unlike the older mathematics-based attempt to understand physical phenomena, insofar as there is no requirement for an adequate mathematical model to be capable of *replicating* the phenomena to be explained: Newton's equations did not produce a new solar system, though they helped to explain and predict observed behaviours in the old one. This paper attempts to explain some of the main features of the design-based approach to understanding natural intelligence, many of them already well known, though not all. The design based approach makes heavy use of what we have learnt about computation since Ada Lovelace. But it should not be restricted to forms of computation that we already understand and which can be implemented on modern computers. We need an open mind as to what sorts of information-processing systems can exist and which varieties were produced by biological evolution.

Keywords:
Design, niche, virtual machinery, physical, mental, model, explanation, biology, AI, commone-sense.

1 How to describe humans and other animals

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 does he/she/it find interesting/boring?
- What is he/she/it attending to?
- What kinds of events surprise him/her/it ?
- 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 treat the whole person or animal (as opposed to functional sub-systems) as the subject of all the verbs (of perceiving, doing, thinking, feeling, deciding, etc.). They make use of a collection of theoretical assumptions and strategies similar to what Dennett called "the intentional stance" and Newell called "the knowledge level" ([4, 14] the differences need not concern us now). That "whole animal" 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 and disastrous emotions, though it is not clear that that would be included in Dennett's "Intentional stance".

There is nothing wrong with such modes of expression if the aim is to entertain, speculate, educate in a general way, generate interest, make excuses, gossip, influence the behaviour of others, or write novels or plays. 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, how they learn and develop, or how their development can go wrong, and if we wish to gain insights into how they evolved, the relationships between evolution and development, and how deliberate external intervention can influence the processes (e.g. educational or therapeutic strategies).

Such scientific understanding is also necessary if we wish to adopt good, reliable, strategies for educating children and

[1] University of Birmingham, UK, email: a.sloman@cs.bham.ac.uk

helping people badly affected by mental abnormalities. Otherwise politicians, educators, therapists and the general public risk being influenced too much by transient fashions and misleading evidence, e.g. evidence based on statistical correlations rather than understanding of mechanisms. (There are "romantic" objections to this approach, based on a dislike of mechanistic explanations, computers, reductionism, or removal of mystery concerning human minds, e.g. [25]. This is not the place to deal with such objections, though they need to be countered.)

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

Going beyond "common sense" descriptions and "correlational" explanations of animal[2] behaviours and competences requires us to formulate theories about the *mechanisms* within the animal that play a role in producing the behaviours, or that produce the competences and dispositions that produce the behaviours, just as explaining how a clock works requires us to identify components which do things that contribute to the clock's functionality, e.g. providing the energy to keep it going, controlling its speed, detecting when to chime, or turn on an alarm sound, etc.

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). 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. Many researchers hope that if only we study physical mechanisms and their connections in brains we shall achieve theories with similar explanatory power. 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).

However brains are far more complex and obscure in their operation than most of the complex systems studied by physicists and chemists or the machines designed (so far) by engineers. It isn't even clear what most of the functioning components of brains 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.[3]

So, on the one hand we get neuroscientists describing components whose ability to produce processes like perceiving, deciding, learning, hypothesising, planning, wanting, evaluating is doubtful and unproven, 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 (a) nobody knows how brain mechanisms could constitute such components and (b) the concepts are too loosely defined for use in a scientific explanatory context (though not for ordinary conversation). The use of such concepts in explanations is often *circular* because the concepts 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 the parts 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., leaving to others the task of specifying more precisely what exactly the parts do and explaining how they do it.

Unfortunately, when AI researchers meet the challenge by trying to specify in constructive terms how the components of intelligent systems (sensory interpreters, memory makers of various kinds, planners, choosers, emotion components, etc.) could be built, they sometimes end up using familiar labels (e.g. "emotion", "perception") for components whose functionality is at best a tiny fragment of what the pre-scientific uses of the labels imply – e.g. robots described as having emotions because they can smile, shake or nod their heads, etc., or described as learning because they change associative weights, or modify rules or databases. McDermott strongly criticised similar tendencies in early AI theorists [8]. The tendency re-emerges with each new wave of fashion in AI.

There have recently been attempts at trying to get beyond this vagueness and circularity by giving the robots physical bodies and sensors or motors that match some features of the biological examples. But that emphasis on embodiment is often too closely tied to attempts to replicate the rather gross morphology of the organisms, both ignoring details like the number of sensors in a mammal's limbs, tongue, lips, etc., and ignoring what humans are able to do if their limbs and some sensors are damaged or missing, like humans born blind, or lacking limbs [22].

3 What sorts of mechanism do we need?

For traditional clocks, all the parts are physically separable parts, whereas we have learnt in the last half century that in the case of complex information-processing systems we need to be able to refer to more abstract concurrently active parts, namely, information-processing sub-systems i.e. pieces of so-called "virtual machinery" that are not physical, though their existence and their operation depends crucially on physical components. There need not be a one-to-one mapping between the VM components and the physical parts. (For more on the complex web of relationships between running virtual machines and the underlying physical machines see [19].)

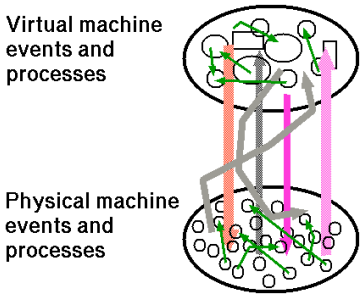


Figure 1. Difference in grain and causal powers between virtual machinery and physical machinery. We also need multiple layers of relative virtuality.

[2] By default read "animal" as including "humans.

[3] See http://news.bbc.co.uk/1/hi/7443534.stm reporting recent research on the previously unsuspected complexity of individual synapses in mammals.

These pieces of virtual machinery are sometimes confused with programs: but programs whether printed on paper or recorded in a computer's memory, are static structures that do nothing, on their own. Something quite different comes into existence when programs run: active, enduring processes in which things get done. For example, games are played, problems are solved, machines are controlled, sensory input is interpreted, things are learnt, goals are considered and selected, plans are constructed, other processes are initiated, suspended, terminated, or in some way modified, e.g. on the basis of *information* about their progress or new developments in the environment.

The connection between processes and programs is not simple, since different processes can use the same program, invoked with different parameters, or running in different contexts – e.g. simultaneously sorting lists of different kinds.

When all that happens, there are very many physical changes in the computer scattered around the memory, the long term file store, registers in the central computer(s), switches in attached interfaces to various external devices, and so on. But the physical processes are not the only processes: things go on in computers (but not *physically inside them* as the transistors are) that are best described in a different language from the language(s) of the physical sciences: e.g. a running virtual machine can consider certain options, investigate some of them, discover that some actions lead to threats and dangers, modify or reorder its goals and preferences, and perhaps find one action that either wins the game or moves closer towards winning. Molecules, rocks, planets, electrical circuits, and, arguably, neurons cannot do those things although processes running on electrical circuits can and presumably processes running on neural, and perhaps sub-neural, mechanisms can also.

AI research has shown how to produce running virtual machines that merit such "high level" descriptions (at least to a first approximation) because of the more detailed descriptions we can give of how they acquire, manipulate, analyse, combine, derive and use various kinds of information and the (internal and external) dispositions they produce and maintain, for example, a disposition to modify a plan to achieve one or more goals in a different way, or to achieve a different goal, if new information turns up during execution.

Some philosophers argue about what is or is not "constitutive" of having an experience or a plan. When we understand, as a result of AI explorations, the space of possible designs and the variety of states and processes that can occur in each design, the importance of much discussed philosophical problems of finding necessary or "constitutive" conditions for something being an example of some pre-theoretic concept becomes utterly insignificant in comparison with the task of exploring that variety, and understanding the implications of the differences. For example, analysing, or experimenting with, working designs to find out which ones can produce systems that have various kinds of experience replaces disputes about *binary* divisions with far more useful analysis of a rich variety of cases.

Understanding diversity in a uniform way is a core goal of biology, and until now Darwinian evolutionary theory (and its more recent ramifications) provides the major example. However a different, but related study of diversity is required for understanding the variety of control systems and the variety of states and processes that can occur in them, including all the things we currently loosely label perception, learning, wanting, experiencing, deciding, disliking, etc.

4 Conceptual tools for explanation designers

Producing such working models depends on, among other things, constructing *information-bearers* of various kinds that express information (for the system using them), and mechanisms for operating on them, for combining, analysing, comparing, storing, searching for and using the information. These information-bearers at the lowest level of virtual machinery in computers are bit-patterns, but at higher levels can be lists of symbols, trees, graphs, arrays, equations, linked collections of changing numerals, sets of rules, running processes, and many more. Moreover these information-bearers, henceforth called "representations",[4] can have different roles in different contexts: the same representation (e.g. a logical expression) could be taken to describe a state of affairs that exists, a state of affairs to be brought about, a state of affairs to be prevented, or a goal to determine whether the state of affairs exists or not. In other words, the operation of virtual machinery in computers is concerned both with semantic content and with control functions of many kinds of representation.

The notion that representations "stand for" or "stand in for" what they represent (e.g. see [24]) is a serious error, since the uses of representations are typically very different from the uses of the things they represent. You cannot climb, or paint, information about the Eiffel Tower. For more on what information is and what representations are, see [20].

At first sight, the chemical and physical structures and processes in brains do not seem to have the power to produce mental states and processes, just as transistors, metallic connections, spinning magnetic discs and the like in themselves do not seem to have the power to play chess, correct spelling, produce goals, beliefs, or plans, nor to take decisions, control inferences, make recommendations, etc. In the latter case (computing systems) we have recently learnt how to give them the powers they seem intrinsically incapable of having (though it is not a simple matter at all to do this: as half a century of research and development has shown – in particular we need far more complex systems than Turing machines). Presumably we shall eventually also learn how the far more complex physical components found in brains (see Note 3) can also be used to support a disparate range of representations and mechanisms, with many different semantic and control functions.

NB: At present, we don't know how neural mechanisms can encode information about complex changing and enduring 3-D structures and relations in the environment, or differences between different kinds of "stuff" that a child learns about, or causal interactions between structures in the environment, or the information used in logical and mathematical processes of reasoning. We also don't know how neural mechanisms represent "meta-semantic" information about things having thoughts, percepts, preferences, emotions, intentions, or information containing gaps. (For more information on the problems, if not the solutions see [23].)

Making progress will require us to develop a deep understanding of the intermediate-level virtual machine functionality that specifies what sorts of information processing is going on in brains, before we can produce good theories as to how the physical components do it. Just looking at the physical and chemical structures and processes found in brains, or even tracing their connections and patterns of activity, can fail to inform us as to what the higher level functions are.

So one important way in which AI may inform biological research

[4] This seems to be the most common use of the word "representation" in scientific contexts.

is by suggesting kinds of intermediate level information-processing functionality that may, one day, suffice to explain the observed competences (and incompetences) in various sorts of animals. Using ideas about such functionality to produce workable explanations at that level may then drive further research into how such mechanisms could be implemented in neural systems, and also research into ways in which they can go wrong, how they develop in individuals, how they differ across species, and how they might have evolved.

Making advances that increase explanatory power of our theories at the intermediate levels is very hard, however – in part, because it is hard to identify the requirements to be met, as decades of over-simplified goals adopted by AI researchers have shown.

5 Whole organism explanations

Often, scientists (or philosophers) attempt to produce explanations of the phenomena observed, or hypothesised, by describing 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. it is you that notices something, not some portion of your brain, or mind, or one of your eyes or ears. What we need are more fine-grained process descriptions, but with the power to explain, and also contradict and refine, our more coarse-grained common sense descriptions. (The corrections would be analogous to corrections of everyday categories in that occurred in previous branches of science: e.g. there isn't a kind of "bad air" that causes illness and whales are not fish.)

Many scientists have tried to avoid those common-sense categories of states, actions and processes, when constructing explanatory theories. Some attempted to use only concepts defined in terms of observable behaviour with consequencs criticised by (among others) Chomsky in [2]. Others, as explained above, try instead to refer to *physical* parts of a person or animal which are assumed to have the required explanatory power. (So-called "mirror neurons" illustrate this move).

A third group refer to *hypothesised* non-physical parts labelled in terms of their cognitive functions, and sometimes represented in diagrams of boxes and arrows, whose operations are supposed to explain what happens. Often there is no specification of how they produce those operations, how to make the components, how to test what they do, or how to vary them, unlike parts of a grandfather clock, which can be made in a factory, given changeable weights or linked in various ways with other parts, and unlike parts of an information-processing system assembled from various hardware and software computing mechanisms that can be combined in different ways to produce different competences.

6 A designer's explanatory requirements

We can build a clock that conforms to a suggested explanation and see if it works like the one that puzzled us – empirical testing. Or we can use our knowledge of geometry, mechanics, dynamics and mathematics to work out what behaviours the proposed structure could produce – analytical/mathematical testing of a theory. But mostly we cannot do either for commonly hypothesised parts of minds, because the specifications are generally too loose – e.g. *being surprised, attending, noticing, wanting, learning, remembering,* etc.

If all we can say to describe hypothesised parts is that they explain the phenomena they were constructed to explain (this bit does seeing, that bit does deciding, another bit does learning, ...) then it is not clear that we are explaining anything.

However, there is another alternative to explanations in terms of categories from behaviourism, physics, chemistry or physiology, that avoids such explanatory vagueness, indicated in previous sections. As a result of decades of research in AI, building on research of other kinds that have steadily extended the powers of computing systems, we have begun to learn how to give explanations of mental processes, and the behaviours they produce, in terms of *information-processing* mechanisms and architectures that are described at a higher level of abstraction than brain mechanisms and at a lower level than common-sense descriptions.

So the vagueness characteristic of many psychological explanations can be reduced, and explanatory power can be increased, by using parts at a lower level of information-processing machinery that we have previously demonstrated in working systems can do the sort of thing we are talking about. This is like going from kinds of matter (water, air, salt, mud, wood, etc.) to kinds of molecules in the history of physics and chemistry. An important feature of this approach is that it assumes that the intermediate level mental mechanisms, states, events and processes are not just useful fictions (as some of Dennett's wording suggests), but actually do *cause* things to happen, including other mental processes and also physical processes, such as changes in brain states and external actions, just as software engineers assume, with good reason, that the calculations and rule applications that go in a computer system can cause both other calculations and rule applications, and also changes in internal physical states as well as on external computer screens and other attached physical devices. This contrasts, for example, with Dennett's "intentional stance" [4] and with "mind-brain identity" theories[5]. For more on the causal states of virtual machinery see [19].

7 Developing a better explanatory ontology

What sorts of parts/components should be referred to in adequate explanations is not easy to understand – and unfortunately there are ill-judged fashions at work. As explained in Section 3, there is a mode of explanation of how complex systems work that is very different *both* from describing their physical, chemical, electrical, or mechanical parts and operations *and* from describing them using common mentalistic language. It is based on what we have learnt about designing complex information processing systems of many kinds, none of which come near the specific kinds of sophistication that we wish to explain in humans (young and old) and other more or less intelligent animals, though there are promising signs of progress, as computer-based systems do more and more things that previously could only be done by humans and other animals.[6] The particular forms of explanation that have been developed refer to such things as

- The kinds of information that the organism acquires and uses.
- How the information is acquired (including which features of the environment make it available and how the sensory and perceptual mechanisms acquire it.

[5] http://plato.stanford.edu/entries/mind-identity/

[6] A particularly impressive robot with a subset of insect-like intelligence is Boston Dynamic's BigDog http://en.wikipedia.org/wiki/BigDog.

- 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 used, from which more complex information is constructable.
- The architectures in which those various capabilities, mechanisms, and information structures are combined.
- How different processes can run concurrently within an architecture, with or without conflicts, and how conflicts can arise, how they can be detected and how they can (in some cases) resolved.
- What sorts of self-monitoring and self-control mechanisms can operate, monitoring and controlling different subsystems.
- How all the items listed above (kinds of information, forms of representation, mechanisms, architectures, etc.) are initially constructed and how 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 (bootstrap) 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 goals, different forms of information-processing, different forms of reproduction, etc.)

In talking about a mind we are talking about a complex system with many concurrently active parts, whose workings need to be explained in terms that can help to bridge the gap between their functions and the underlying physical mechanisms that make it possible to have such functions and which limit and shape those functions. We also need an explanation of how those parts interact, harmoniously most of the time, and what happens when they come into conflict.

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 acquire a deep scientific understanding (as opposed to close human-to-human empathetic understanding for example):

- Which parts of the information-processing architecture are involved in the capability or deficiency?
- 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 sub-systems 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,...)?

For example, a parrot can use one foot to balance on a perch while at the same time alternately holding a walnut in its beak or the other foot, as it rotates it trying to find a good place to bite into it. A particular robot could be built that does nothing but produce that behaviour, though only with a very specific shape of perch and a very limited range of sizes and shapes of walnut. But a robot built using contemporary AI techniques would not be able to *work out* how to do such things starting from a reportoire of competences that could be used for a wide variety of purposes, e.g. alternately using beak and feet while climbing. (Some of the problems, but not all, are due to limitations of current mechanical and electronic engineering, including power weight ratios, strength weight ratios, sensor limitations, etc.)

8 Limitations and common errors in current AI

Of course, where current AI researchers and their models use flawed assumptions, theories or designs, biologists need to resist being inspired by them! For example, the switch from common-sense descriptions to descriptions that have been useful in designing working computational models is not always an improvement, since sometimes the ontology available to AI 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 mostly 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 [18].

An issue that some AI designers do not face up to is whether to think of all software components of an intelligent system as running on a single processor, or as using multiple concurrently active processors. Very often the only development environment available to AI researchers assumes either one or a very small number of processors, and uses programming technology that does not easily support development of asynchronous concurrently active subsystems. A result of this is that such researchers have to work on techiques for interleaving different activities, and strategies for optimising the allocation of processing resources between them. An example is a great deal of work involving addition of metacognitive capabilities to intelligent systems, which has addressed the need to optimise the sharing of resources between cognitive and metacognitive processes, and possibly other processes.

All that effort is wasted in the context of modelling natural cognition, if brains have different subsystems running in parallel performing the different tasks. An extreme version of the error

is the widely accepted assumption that intelligent robots need to make use of a repeated cycle of cognitive phases such as this:

sense→think→decide→act

In contrast, our own work has assumed that many processing components of biologically inspired robot can, and need to, run in parallel, as indicated in our work on the CogAff architecture scheme and its H-Cogaff special case (http://www.cs.bham.ac.uk/research/projects/cogaff/#schema).

Also bad theories in AI, e.g.

Symbol grounding theory

theories about emotions

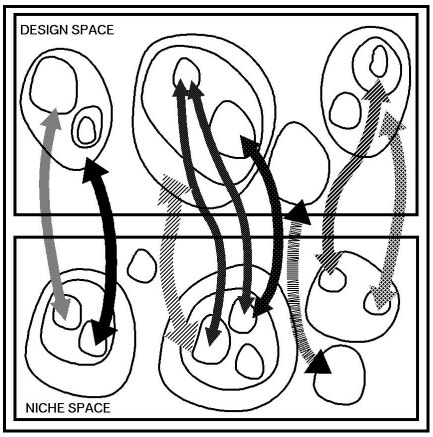
theories that all motivation has to be reward based.

Another restrictive influence can be the assumption that existing information processing architectures or other existing techniques are adequate to the task of building explanatory models, when what is badly needed is new deep requirements analysis: examining the biological phenomena (e.g. animal behaviours, or studies of brain mechanisms) with very great care to see whether something can be learnt that improves the way in which AI can serve biology.

9 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 costs and benefits of the *actual designs* found in humans.



We need to understand different sets of requirements (niches) different designs, and the many complex ways designs can relate to requirements (not just in terms of numerical fitness functions). We also need to understand the many kinds of trajectories through niche space and design space that can occur in evolution, in individual development, in cultural evolution, and so on. Dennett's little book on "Kinds of minds" [5] illustrates this, but using very "broad brush" categories.

10 Kinds of requirements

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.

- 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.
- 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 competences**, which not all organisms seem to have.
- Being able to use "self-directed" meta-semantic competences applied to one's own thinking, reasoning, perceiving, etc., for instance finding a flaw in one's planning strategy and repairing it. Can it recurse?

11 Kinds of Observations Needed

A task on which more thought is required is how the research goals listed here can influence choice of biological experiments and the observations required. (See the paper by Chappell and Thorpe in these proceedings.)

One implication that is rarely noticed 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! (I see many presentations where graphs show changes in averages without showing any of the variance, alas, indicating a deep flaw in our education of scientists.)

12 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 now calls 'the designer stance', in [7]) 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.[7]

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 much richer spaces of possibilities, including *taxonomies* or *generative grammars*. We should not assume that the only alternative to a binary division is a linear continuum (differences of degree). Biological changes, insofar as they are based on molecular changes and other structural changes (e.g. duplication of a component) are inherently *discontinuous*. Our theories of the spaces involved must accommodate this.

13 An example, from AI

Many people (including the author years ago) 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 documented them in [21]. I don't claim that the analysis of that document is complete: there are important sub-cases to be distinguished, especially if we are to understand the stages in the evolution of the more complex competences.

When considering any cmpetence or mechanism of type 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?

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

[7] As explained in this talk on "Why robot designers need to be philosophers – and vice versa"
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk51

This idea was used in an analysis of the notion of 'free will' originally posted to a 'usenet' news group, now available in [15]. Similar ideas are in [3] and [6].

14 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 [19] that include concurrently active interacting sub-systems, including some sub-systems 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.

15 Logical geography vs logical topography

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. One of the papers, [16]. 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 constituting 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, we can see that philosophical theory building has much in common with scientific theory-building (including the ability to introduce extensions to our ontology), and which uses abduction. However, for philosophers and researchers studying natural intelligence, to ignore the advances in the logical topography of information processing systems would be analogous to chemists continuing, like the old alchemists, to find and use laws of how different kinds of stuff interact without attempting to move to a deeper level of explanation, as emerged in the atomic and sub-atomic theory of matter.

16 Conclusion?

There is no conclusion, because we are still in the early stages of a very rich and deep exploration whose implications are potentially profoundly important not only for the biological sciences, but also for our understanding of ourselves. A revised, extended version of this paper will be written after the symposium, possibly with co-authored sections.

ACKNOWLEDGEMENTS

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 his paper [7]. I tried to take his ideas a bit further in [17] and the presentations here: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/ Much of Minsky's work, is in a similar vein. I think he has been moving upwards in the layers of virtual machinery in his recent work, and it is not always easy for readers to understand how the ideas might be implemented. See [9, 10, 11, 12, 13] The most exhaustive and historically deep survey of these ideas can be found in Margaret Boden's "magnum opus", [1]. Two extracts from that have been made available for this symposium.

REFERENCES

[1] M. A. Boden, *Mind As Machine: A history of Cognitive Science (Vols 1–2)*, Oxford University Press, Oxford, 2006.
[2] N. Chomksy, 'Review of skinner's *Verbal Behaviour*', *Language*, **35**, 26–58, (1959).
[3] D.C. Dennett, *Elbow Room: the varieties of free will worth wanting*, Oxford: The Clarendon Press, 1984.
[4] D.C. Dennett, *The Intentional Stance*, MIT Press, Cambridge, MA, 1987.
[5] D.C. Dennett, *Kinds of minds: towards an understanding of consciousness*, Weidenfeld and Nicholson, London, 1996.
[6] Stan Franklin, *Artificial Minds*, Bradford Books, MIT Press, Cambridge, MA, 1995.
[7] J. McCarthy, 'The well-designed child', *Artificial Intelligence*, **172**(18), 2003–2014, (2008). http://www-formal.stanford.edu/jmc/child.html.
[8] D. McDermott, 'Artificial intelligence meets natural stupidity', in *Mind Design*, ed., J. Haugeland, MIT Press, Cambridge, MA, (1981).
[9] M. L. Minsky, 'Steps towards artificial intelligence', in *Computers and Thought*, eds., E.A. Feigenbaum and J. Feldman, 406–450, McGraw-Hill, New York, (1963).
[10] M. L. Minsky, 'Matter Mind and Models', in *Semantic Information Processing*, ed., M. L. Minsky, MIT Press, Cambridge, MA, (1968).
[11] M. L. Minsky, *The Society of Mind*, William Heinemann Ltd., London, 1987.
[12] M. L. Minsky, *The Emotion Machine*, Pantheon, New York, 2006.
[13] M.Minsky, 'Interior Grounding, Reflection, and Self-Consciousness', in *Brain, Mind and Society, Proceedings of an International Conference on Brain, Mind and Society*, Tohoku University, Japan, (September 2005). Graduate School of Information Sciences, Brain, Mind and Society. http://web.media.mit.edu/~minsky/papers/Internal Grounding.html.
[14] A. Newell, 'The knowledge level', *Artificial Intelligence*, **18**(1), 87–127, (1982).
[15] A. Sloman, 'How to Dispose of the Free-Will Issue', *AISB Quarterly*, **82**, 31–32, (1992). http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#8.
[16] A. Sloman, 'Two Notions Contrasted: 'Logical Geography' and 'Logical Topography' (Variations on a theme by Gilbert Ryle: The logical topography of 'Logical Geography'.)', Technical Report COSY-DP-0703, School of Computer Science, University of Birmingham, Birmingham, UK, (2007). http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0703.
[17] A. Sloman, 'The Well-Designed Young Mathematician', *Artificial Intelligence*, **172**(18), 2015–2034, (2008). http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0807.
[18] A. Sloman. Ontologies for baby animals and robots. From "baby stuff" to the world of adult science: Developmental AI from a Kantian viewpoint., 2009. Online tutorial presentation: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#brown.
[19] A. Sloman. Virtual Machines and the Metaphysics of Science, Sep 2009. http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#mos09, PDF presentation for AHRC Metaphysics of Science Conference.
[20] A. Sloman, 'What's information, for an organism or intelligent machine? How can a machine or organism mean?', in *Information and Computation*, eds., G. Dodig-Crnkovic and M. Burgin, World Scientific, New Jersey, (To appear). http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#905.
[21] Aaron Sloman, 'Requirements for a Fully Deliberative Architecture (Or component of an architecture)', Research Note COSY-DP-0604, School of Computer Science, University of Birmingham, Birmingham, UK, (May 2006). http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604.
[22] Aaron Sloman, 'Some Requirements for Human-like Robots: Why the recent over-emphasis on embodiment has held up progress', in *Creating Brain-like Intelligence*, eds., B. Sendhoff, E. Koerner, O. Sporns, H. Ritter, and K. Doya, 248–277, Springer-Verlag, Berlin, (2009). http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0804.
[23] A. Trehub, *The Cognitive Brain*, MIT Press, Cambridge, MA, 1991. http://www.people.umass.edu/trehub/.
[24] B. Webb, 'Transformation, encoding and representation.', *Current Biology*, **16**(6), R184–R185, (2006).
[25] J. Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, W.H.Freeman, San Francisco, 1976.