(Last updated 30 May 2000)

LETTER TO DAN DENNETT

Hi Dan,

I've recently posted you a couple of papers, one over a year old but published in March, and another on architecture-based conceptions of mind, to be published later in proc 11th int. Congress on Logic Methodology and Philosophy of Science.

If you are too busy, maybe some of your students will find something of interest in them.

There's lots more in the Birmingham Cognition and Affect directory, http://www.cs.bham.ac.uk/research/cogaff/ including a recent discussion note arguing that AI has nothing to do with Turing machines, except in the minds of some people who attack AI.

I sent you those papers because I've been reading parts of Brainchildren and, as usual, I very much like most of what you've written but feel uncomfortable about other parts; yet I feel that on the basis of the areas of agreement it really ought to be possible for the us to reach agreement and possibly even produce some sort of manifesto based on it!

Below I have tried to identify a list of important points on which I think we ought to agree: some of which you'll readily accept, and some you may not. It would be interesting to try to resolve the disagreements on the latter, since I believe that the views I hold on those points are either derivable from or at least better integrated with the views on which we agree than the opposing views are.

Incidentally I nearly always agree with you when you are attacking other philosophers. E.g. I completely agree that most of those who propose zombie-based theories do not realise that they have not specified their zombies in sufficient detail to be relevant as counter-examples.

When specified in sufficient detail, the claim that the zombies lack something we have becomes either totally implausible or incoherent. (Usually the latter.)

However, I'll try to indicate below that your attack on Fodor is partly mistaken (though I think he is wrong about many things!)

The main thing I like, both in the papers I have so far read in Brainchildren and also in the Kinds of Minds book, is that you seem to have moved more and more into "design mode".

You probably don't recall my contribution to your BBS treatment on The Intentional Stance around 1988. My little piece was entitled "Why philosophers should be designers". I got the impression from your comments that you had brushed it aside because you thought I was arguing that people like you should give up philosophy and take up programming. I wasn't. I was recommending what you have actually been doing in the last 10 years or so!

Your recent work has illustrated these two points quite well:

(a) One can be a designer and a philosopher simultaneously. (Kant was! I see our work as a sort of continuation of what I think he was trying to do, though he lacked conceptual and other tools for the task).

(b) Doing design work does not necessarily involve writing programs.

I do all three (philosophy, design and programming), but my programs are mainly low level tools for my students and collaborators use, i.e. an agent toolkit -- see:

http://www.cs.bham.ac.uk/~axs/cog_affect/sim_agent.html
http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

My own designs (and yours too) are still too abstract, complex, vague, under-specified etc. to be directly implementable, though I hope eventually to make a lot of progress with that, if I live long enough!

I am learning a lot from interacting with neuroscientists, and also from trying to use evolvability as a design constraint, not because I think evolvability is a requirement for all interesting types of mental architectures, but because that constraint gives useful insights into our own virtual machine architectures. (The plural is deliberate.)

Some people require something like metabolism in the underlying implementation, but I think that elevates an empirical fact to a conceptual necessity (more on that below.)

Leaving aside minor differences in the sorts of designs we propose, and our styles of presentation, and argument, I think the main remaining point of disagreement between us is that you still have some attachment to the intentional stance as the basis for ascribing semantic capabilties, whereas I have always thought it is redundant, if the design stance is developed properly.

I would like to try to diagnose that disagreement and possibly remove it, perhaps by coming up with a specification for a slightly different set of stances on which we'd agree. Or maybe I too have misunderstood what you say about the design stance...

As a step towards this clarification (and unification?) I've tried to produce (without being terribly scholarly and giving full references, etc.) a summary of a collection of conjectures, etc. many of which I think you share, some of which I suspect you share though I am not sure, and others which seem to contradict things you have claimed.

I think the points we agree on are incompatible with the intentional stance explanation of how we ascribe mental states. Or rather, they appear to me to make the intentional stance redundant for that purpose.

I'll start with one point I am fairly certain you will agree with. Then move through a succession of claims where I am less sure about your view. I'll try to indicate where I expect you will disagree, as an indication of possible misunderstandings of your position. I apologise for the length of all this. (I thought it would be much shorter when I started writing it.)

THESES/CONJECTURES

1. If I ask whether a prolog or lisp virtual machine is currently running in a certain computer, that is

a question of fact.

This is not a question of fact about its *externally* observable input/output behaviours, since the VM may be running a complex problem-solving or simulation program which will run for a very long time, or even forever, and need never print anything out.

(Alternatively it may be modelling, or implementing, a disembodied pure mathematician who never communicates with anyone else, but spends all his life thinking about hitherto unsolved mathematical problems, exploring clues, refuting conjectures, analysing failed proofs, etc. etc.)

(1.a. (Digression with which you may disagree). A corollary of (1.) is that versions of functionalism that emphasise the necessity of inputs and outputs, i.e. interactions with a physical environment, are mistaken. There could be nothing but a large collection of internal (virtual machine) processes interacting with one another.)

2. Although it is a question of fact whether a Lisp VM is running on a particular physical computer, it will in general be quite difficult to check out the answer, especially if we don't know anything about the computer or the programs running on it.

However, if it is a familiar make of computer and I know which operating system is running on it, and I have designed and run the program using a well-known interpreter for Lisp (say), then I can be pretty confident that the Lisp VM is running.

(When I use a *compiler* rather than an *interpreter* it is not so clear whether a Lisp VM is running or something else which is partly behaviourally equivalent. I have argued that these are importantly different cases, though not all computer scientists agree with me.)

3. To say that a particular VM is running is to say that there is an information processing machine with a certain architecture, i.e. various co-existing parts which interact causally, e.g. many arithmetic routines, list processing routines, storage allocator, garbage collector, procedure control stack, various mechanisms for handling temporary information about what's going on e.g. for profiling or tracing purposes, etc., and also user programs, which may themselves extend the architecture in a quite complex way.

It's worth noting that some programs running on the lisp VM will produce an algorithmic sequential VM process. Others will produce a collection of concurrent interacting processes with enduring states. Interesting models of mind will generally be of the latter kind.

(For the concurrent processes to interact asynchronously the VM needs some deep causal connection with things outside the VM, e.g. a clock in the machine, or physical transducers capable of triggering the Lisp interrupt handler.)

4. Events in the VM can cause both other VM events and also physical events to occur, and vice versa. I.e. VM events are not epiphenomenal.

None of this presupposes that there are any causal gaps in the physical machine: causal completeness at the physical level does not rule out non-physical causes of physical events.

5. That's because causation is not a kind of conserved fluid, but rather a matter of which of a complex collection of conditional statements, including many counterfactual conditionals, is true.

This is how we normally think about causation, contrary to what some philosophers assume. E.g. we believe ignorance can cause poverty, poverty can cause crime, the spread of knowledge can cause technological change which produces massive physical changes (e.g. new buildings, bridges, and even global warming). None of this presupposes causal gaps in physics. Physics is an implementation machine for socio-economic virtual machines. (I expect you still agree? Is that right??)

6. All biological organisms are physical and chemical machines which include information processing machines (VMs) implemented in those physical and chemical machines (possibly using intermediate level VMs, such as neural machines).

These virtual machines vary enormously in their architectures and we currently know very little about the specific VM architectures of most types of organisms, though we can do a lot of well informed speculation about this (as you and I and others do).

In particular, biological systems typically involve a huge number of homeostatic virtual machines operating concurrently at many different levels of abstraction. Some of this activity may engage with processes in the high level VM, e.g. triggering high level goals, evaluationts, redirection of attention, etc. (More on this below.)

7. Some biological VM architectures may involve several levels of virtual machines implemented in virtual machines, etc. (This is also true of computers: a Sparc or pentium is a virtual machine implemented in digital circuitry, and may be implemented in different ways at different times, e.g. as computer technology advances.) Layered VMs are even harder to investigate empirically than single-level VMs.

It is often assumed that everything in the human VM architecture is implemented in a connectionist VM. But it could be that more than we currently know is implemented in chemical engines which run in parallel with and partly implement the neural engines. (E.g. mood modulation.)

8. Finding out what the information-processing VM architectures of various kinds of organisms, including chickens, eagles, rats, chimps, adult humans, newborn infants, etc., are is a *factual*, scientific task (which can benefit from work in many disciplines, including philosophy, linguistics, psychology, anthropology, neuroscience, chemistry, physics, biology, ethology, etc.)

Although it is a factual task, we don't necessarily already have all the right concepts for doing the task: our ideas about possible architectures, mechanisms, types of interaction will have to be clarified and extended, as has been happening steadily since we started developing, programming and using computers.

(Quine's arguments about radical indeterminacy of mental states seem to me to ignore the possibility that some of the indeterminacy is resolved by the precise details of the implementation. As far as I know he never considered that option because he is essentially a kind of behaviourist, analysing mental states in terms of global input-output relations rather than in terms of states of the internal, unobservable, virtual machine architecture. I am not sure how much you agree with Quine on this.)

9. Although the physical implementation in some sense *determines* which VM is implemented, I don't believe that we can start from a description in the language of physics and *derive* a description at the level of a Lisp machine, or chess machine, or a monkey's mental architecture, using purely logical or mathematical reasoning.

The main reason for this is that the VM architecture specification uses concepts that cannot be *defined* in terms of the concepts of physics. The VM ontology is definitionally irreducable, but implementable none the less.

(Chalmers got this wrong in his discussion of supervenience. There's lots of hard work still to be done clarifying this. Matthias Scheutz did some interesting relevant work for his PhD at Indiana, supervised by Brian Smith. He is coming to work with me from June, so maybe we'll sort that out!!)

I am a bit like Searle: I say that the physical architecture of a brain can *cause* the VM architecture of a mind to exist. But in most respects I disagree with him: the same VM can often be caused in very different ways. In fact we are still largely ignorant about which VMs can be implemented in which lower level systems. More on this below.

10. (You may not like this). Many information processing systems include semantic capabilities: parts which refer to other parts, describe other parts, store information about what has happened in other parts, make predictions about what will happen in other parts, produce instructions that control what happens in other parts.

There can also be reference to items in the environment, e.g. which network connections are up, which down, which other machines are currently transmitting information, how many packets of a message have so far been transmitted, what the ambient temperature is (if it gets too high some machines will send email to an operator and then switch themselves off).

In biological organisms, the external semantic capabilities include detection of opportunities, dangers, various types of affordances, etc.

These semantic capabilities are implemented in mechanisms which do not have semantic properties. The semantic capabilities arise out of the ways the various components of the virtual machine work: what they do, and what the effects are. Negative feedback loops are a very simple example, but there are far more complex and sophisticated examples.

E.g. in a lisp VM the memory management subsystem must have information about how much unused heap space there is. When a lisp instruction to create a new data-structure is executed the system must be able to work out whether there is enough spare space for that data-structure. If not it will initiate a garbage collection process, which also requires using a lot of information about what is where in the heap, and what needs to be changed if structures are relocated by compaction, etc.

(It is often assumed that the basic type of semantic relation between atomic symbols requires a causal connection with systematic co-variance of some kind. However that is false, not least because it rules out reference to non-existent entities, and also one-off symbols as in "Let this thimble represent Napoleon, that box Moscow, etc." In a computer a bit pattern can refer to a non-existent memory location, not because of any causal correlation but because of how the bit pattern is used by the system in fetching, storing, comparing information, etc. The semantic link is set up by the mechanisms that use it. Reference to a non-existent memory location will typically be trapped by the hardware and generate an error interrupt. Or it may be trapped earlier by software. But there's still reference to a non-existent location. Reference to locations in virtual memory is more subtle and complex.)

In at least some cases whether a machine includes these semantic capabilities is a question of fact. But it is not necessary for the mechanisms providing the capabilities to be optimal, or particularly good as long as they work.

11. Whether a machine (artefact or organism) does or does not have and use semantic capabilities is a factual question.

However "semantic capability" is a *cluster* concept. When we attribute such capabilities to an organism or machine, there are many different combinations of capabilities that may or may not be present, but we do not at present have a clearly defined concept of "semantics" which determines precisely which combinations are and which are not sufficient. (It's something like these other cluster concepts: alive, conscious, intelligent, animal, machine, etc.)

By exploring the space of possible architectures and niche space (the space of possible functional requirements for such designs) in more detail we'll be able to come up with many new refined versions of this cluster concept (semantic capability, or information processing), and then it will be much clearer which concepts apply to which designs -- though indeterminate cases will always remain.

(That's a bit like getting clearer about elements, isotopes, compounds, mixtures, alloys, etc. as a result of advances in our understanding of the architecture of matter.)

NB: People sometimes wrongly assume that we are dealing with a continuum of cases, and decisions about where to draw boundaries are totally arbitrary.

The space of designs is NOT a mathematical continuum. On the contrary, it has very many natural discontinuities, most of which we have not yet noticed, or analysed, or taken into account in developing our concepts.

It is often forgotten that Darwinian evolution is inherently DIScontinuous. Once it is clear that the changes have to be discontinuous it should not be surprising that some steps are small and some large. (Non-mathematicians often confuse "gradual" and "continuous". I think you explain this in Darwin's Dangerous Idea.)

So there are lots of distinct concepts waiting to be discovered and used in classifying different sorts of machines. Questions we ask now, and hypotheses we formulate now, will almost certainly be extended or replaced by more precise versions that we'll understand in years to come.

E.g. the kind of semantic capability a single-celled organism has when it detects and reacts to an affordance in its environment is very different from that of a chimp or human, but we don't yet know in how many ways it is different.

We don't yet have a good "periodic table" of semantic capabilities into which we can fit these and other organisms and artefacts.

[Similar remarks can be made about many other cluster concepts used in describing mental states and processes.

Do you remember who first introduced the notion of a "cluster concept"? I think I picked it up when I was a DPhil philosophy student in Oxford 1957-62 but I cannot remember from where. It's related to Waismann's notion of "open texture" and Wittgenstein's "family resemblance concepts". It is not the same as "vague concept".]

12. (You may not like this). None of the semantic capabilities mentioned above presuppose that the system is in any way *rational* (though in some cases their human designers are!).

They are simply mechanisms which have various capabilities and when they work normally they bring about various results, some good, some bad for the system (or for something else which uses the system).

They can work abnormally also. E.g. when information in the Lisp VM's heap is corrupted, the mechanisms will not necessarily notice the corruption, and the result of a garbage collection when the heap is corrupted may be to cause some datastructures to be over-written with junk. So mechanisms with semantic capabilities can be far from optimal, and even distinctly buggy.

13. (You may not like this) Most of the semantic capabilities of most organisms have nothing to do with rationality, in the sense of using beliefs and goals and working out the best way to achieve certain goals.

In fact, the vast majority of organisms (e.g. insects and simpler organisms) are not capable of deliberation, and therefore cannot do it rationally or irrationally.

Even organisms with a purely reactive architecture can acquire, store, manipulate, and use information.

Examples are: detecting things, using neural nets or other mechanisms to learn useful generalisations, using information to guide internal or external behaviour, selecting options, etc.

None of that requires the VM to be the least bit rational (though one might regard the evolutionary processes as functioning partly like a rational designer).

The mechanisms have semantic capabilities and they use them, but not in any rational way. When some bits go wrong, other bits may continue doing what they always do, even if it's a complete disaster for the system as a whole (like garbage collection overwriting datastructures).

I think a big mistake Newell made in his discussion of "the knowledge level" was to assume that rationality was required for semantic capabilities. His collection of levels did not have enough intermediate levels. (I may be misremembering what he wrote.)

14. (You may not like this) For some kinds of (virtual) machines, when we investigate their semantic capabilities it suffices to adopt the design stance: we don't need the intentional stance.

I.e. notions like "refers", "describes", "derives", "calculates", are part of the the required ontology of a VM designer. Such a designer need not adopt the intentional stance, or attribute rationality, in order to justify the use of such concepts.

These concepts are commonplace tools for software engineers. You cannot design a good operating system, compiler, office automation system, factory control system, etc. without making use of these notions. Even the lower level designers of general purpose Von Neumann computers have to talk about addresses (pointers) and instructions. These are part of the specification of the Sparc or Pentium or Dec Alpha VM.

Likewise when we analyse an animal's visual system, to find out how it segments information in the optic array, how it interprets 2-D patterns in terms of 3-D structures, how it classifies objects and relationships, how it uses that information to trigger learning, decision making, or alarm reactions, etc. we are assuming semantic capabilities, but not assuming rationality. We adopt the software/hardware engineer's stance, not the intentional stance.

Conjecture:
I have a hunch that when you first introduced the distinction between the physical, design and intentional stances, you thought of design as being concerned with physical organisation and physical interactions.

Since semantic relationships cannot occur there, you thought that only the intentional stance could support them. By making a distinction between a physical design stance (used by a designer of a bicycle or old fashioned typewriter) and a software (or virtual machine) design stance, you could have incorporated semantics (and therefore at least simple types of intentionality) into the (software/VM) design stance.

That would have reduced the interest and importance of the intentional stance.

Of course: it's not totally simple since we still don't have a very clear idea how to justify talk of software virtual machines: we just know that it is possible, and necessary, and it works well for engineers of many kinds, though most biologists and brain scientists don't know about it, or are only just beginning to learn about it.

Many philosophers think that if they know what a turing machine is they they know everything about virtual machines. They are just wrong.

15. (I expect you agree with this though I am not sure: it's part of the design stance). Just as learning about the architecture of matter enabled us to get clearer about many of our concepts of kinds of physical stuff and kinds of processes involving physical stuff, similarly learning about the VM architecture of a mind can give us a basis for specifying a host of mental concepts relevant to states and processes that can occur in that mind.

Minds with some sort of rationality will be a special case. Primitive rationality is provided by a deliberative layer in the architecture along with mechanisms for generating goals, comparing them, adopting or rejecting them, and deriving plans from them.

More sophisticated forms of rationality require a meta-management (reflective) layer, providing some form of self-consciousness, including evaluation of internal processes, states, strategies, motives, etc.

(The distinction between deliberative mechanisms and meta-management mechanisms is being elaborated as part of the Birmingham Cognition and Affect project. E.g. it leads to several different concepts of emotion. The distinction can be found in Minsky's work, though it is more prominent in his new -- unpublished -- book.)

16. (You may like some of this, but not all?) We are innately disposed to be software engineers.

We don't come to believe that there are other minds by making any of the kinds of inferences that have been proposed by philosophers. We are "biologically programmed" to assume there are other minds because it is of biological use to do so.

E.g. we don't reason "I have these inner states, and I have various kinds of visible properties, therefore other people who have similar visible properties must have similar inner states". We don’t *define* mental concepts in terms of patterns of behaviours or input-output relationships. We don’t explicitly adopt the intentional stance and assume that others are rational. (Some philosophers may do that, but young children don’t need to.)

Likewise we don’t come to believe in the existence of 3-D space and 3-D spatial objects as a result of making inferences from patterns in 2-D sense data, as some epistemologists have suggested.

Rather we are born with genetic pre-dispositions to build an ontology which includes a 3-D spatio-temporal environment with many kinds of objects, properties, relations, events and processes within it, including other agents with perceptual abilities, beliefs, desires, intentions, and states such as anger, approval, etc.

Similarly many engineers now design systems with built in ontologies that enable them to interact with a part of the environment.

Biological evolution, not philosophy, solves the "other minds" problem, by designing us to function in an environment containing other intelligent organisms.

Different organisms use different ontologies, matched to their needs, their capabilities and the structure of their perceptual mechanisms. In some cases the ontology is largely determined innately (e.g. newly hatched chicks and other precocial species can walk about and peck at grains of food) whereas other organisms have to develop some aspects of the ontology during infancy.

In humans (and other altricial species) the information processing architecture is not fully formed at birth, but is built up in the early years while parts of our brains and their interconnections grow under the influence of complex processes of interaction with the physical and social environment.

During that time aspects of the ontology used in perceiving and acting in the environment will be constructed as a result of interactions between innate predispositions and what is encountered in the environment. (Language learning is a special case of this.)

Some of this can include attributions of mental states with semantic content.

A hunted animal may need to be able not only to recognise its hunter but also to work out which way the hunter is facing etc. A social animal may need to be able to tell whether another animal is looking at it, whether it is angry, threatening, etc.

I.e. a mentalistic ontology is not a philosophical invention but a biological necessity. (Maybe that’s what Strawson was trying to say in 1959 when he said we use M predicates and P predicates?)

The conjecture is that humans naturally develop ways of thinking of and *perceiving* humans and other animals, and themselves, using a presumed information processing architecture. Seeing someone as happy, sad, as looking in a certain direction, as trying to do something and failing, as threatening, as a friend or as a foe, etc. uses perceptual mechanisms which map low level perceptual data produced by physical transducers into abstract perceptual categories.

This is no different in principle from the perception of other abstract "affordances", e.g. edibility, graspability, obstruction, rigidity, flexibility, strength, support, etc. (I argued this at more length in Sloman 1989, criticising David Marr’s one-way, data-driven, geometry and physics-based model of visual perception).

The mentalistic ontology may be implicit in parts of the reactive architecture. For organisms which don't merely perceive and react, but plan, form intentions, and do 'what if' reasoning about others, it is likely that they will find a more *explicit* mentalistic ontology useful.

This ontology could refer to a presumed virtual machine architecture which is assumed to exist in other agents.

In other words, without learning any philosophy, and without being aware of what we are doing, we act as designers, as software engineers. (My "Architectural requirements..." paper argues that this is implicit in the thinking of novelists, playwrights, poets, and gossips.)

It could be that this "natural" mentalistic ontology presupposes rationality. I think that's unlikely.

Alternatively it could be that it presupposes only a variety of vaguely specified information processing mechanisms which support various capabilities: acquiring information, storing it, transforming it, generalising it, triggering goals, comparing goals, associating goals with actions, activating actions, forming new plans, etc. Some of these capabilities could be reactive, others deliberative. Only certain deliberative capabilities require any sort of rationality.

For example, when we assume that some people can sometimes get into states in which they are easily irritated, we are assuming that there are mechanisms that produce this result, but we don't need to assume that there's anything rational about it. When we assume that alcohol, or jealousy, can modify perceptual capabilities, motor control capabilities or decision-making capabilities, we don't need to adopt the intentional stance; we just assume that the capabilities and behaviours of some semantically informed components of the virtual machine get changed.

I.e. we are thinking as designers trying to identify and perhaps explain bugs in the functioning of an instance of a design.

I believe that all humans older than about 1 year (and maybe some other animals) can do this to some extent, but people vary enormously in the richness of their grasp of the underlying architecture, just as they vary in their grasp of the physical properties of things in their environment.

(Likewise: a woodworker, a dressmaker, a gardener, a painter, and a sculptor, will understand different things about the properties of physical materials.)

People also vary enormous in their ability to articulate what they grasp: and most of us cannot articulate more than a small portion of it.

(Shakespeare was particularly good at certain aspects of it. Kant got some parts right, but most of it wrong. The British empiricist philosophers got most of it wrong. Philosophers tend to be too influenced by bad theories produced by previous philosophers. Psychologists tend to be too frightened to talk about things they can't observe or measure.)

17. (You probably agree with this) A corollary of all the above is that science extends and refines but does not eliminate or replace our pre-theoretic mental ontology.

As we develop more accurate, detailed and precise theories of the types of information processing architectures making up various kinds of human minds, and use those architectures as a basis for clarifying concepts of the kinds of mental states and processes that the architecture can support, we shall find that our pre-theoretical assumptions were not completely wrong, and that the concepts were not completely useless.

Rather we shall observe a type of extension, refinement and clarification which is partly similar to what happened to our concepts of kinds of stuff, and of kinds of physical processes, as we learnt more about the architecture of matter.

(The explicit philosophical and scientific theories we formulate will probably have to be rejected: but I am not talking about those.)

Thus the position summarised here is inconsistent with eliminative materialism, or extreme forms of anti-folk-psychology connectionism.

18. (I think this bit contradicts some of what you have written, e.g. in attacking Fodor in Chapter 4 of Brainchildren). An architecture including a deliberative layer can use explicit propositions.

A robot's or organism's information processing architecture may have different sub-architectures that function in different ways, and some aspects of the sub-architecture will map more closely onto pre-theoretic ways of thinking than others. Likewise some theories will be correct if applied to one part of the architecture and wrong if applied to other parts.

E.g. GOFAI models of belief and knowledge are more suited to deliberative sub-architectures, connectionist models more suited to reactive sub-architectures (even if connectionist mechanisms are used by the brain to *implement* the deliberative mechanisms.)

In our work on the Birmingham Cogaff Architecture, partially developed in the papers listed below, we assume that there are reactive mechanisms, deliberative mechanisms and meta-management mechanisms, which evolved at different times. (This is partly similar to ideas in Dennett 1996.)

The pre-theoretical concept of believing that something is the case, or knowing something, maps into these different sub-architectures in very different ways.

For instance, reactive mechanisms may include many generalisations e.g. about what will happen next in various contexts, or which actions should be performed when, all *implicitly* encoded in collections of weights in neural nets or other mechanisms. In fact the same generalisation may be encoded in different ways in different parts of a reactive mechanism, if different parts are used for different purposes or in different contexts.

E.g. much knowledge about gravity and the physical properties of matter is encoded in parts of the system used for running, walking, throwing things, catching things, lifting things, etc. Such information about effects of gravity and how to do things in a gravitational environment may be encoded in quite different ways in different parts of the reactive architecture involved in these different tasks.

In humans, much of the knowledge embodied in expertise produced by long training is encoded entirely in the reactive subsystem, such as the knowledge used by an expert tennis player to anticipate an opponent's move. (That's why an expert flautist may be unable to tell you the fingering required for top F-sharp, but can show you which fingers are up and which are down, while holding a flute.)

By contrast, the store of information used by a deliberative component may have many explicitly articulated propositions such as "Arsenic is a poison", "Americans drive on the right hand side of the road", "Aspirin can relieve headaches", "Triangles form rigid structures, "Squares do not form rigid structures" (both of which I learnt as a child from playing with meccano), and many algebraic and trigonometric equations, and other mathematical theorems and scientific laws, e.g.

$$a^2 - b^2 = (a + b) x (a - b)$$

Mathematics and rule-governed board-games would be impossible otherwise.

There are different kinds of "implicit" information. The information implicit in weights in a neural net, or patterns of activation of a collection of neurons, is one kind.

By contrast, in a deliberative mechanism with rich deductive capabilities, information may be *deductively* implicit, i.e. implicit because it is inferred when needed and only then made explicit.

Much negative information will be implicit in that sense. E.g. you can ask me many questions about the contents of my house, to which the answer will be No (e.g. is there a swimming pool in the kitchen?), but not because I already have the negated proposition in my collection of explicit beliefs.

Likewise, if there is a meta-management system some of its self-monitoring, self-assessments, and self-controlling may be done implicitly using reactive mechanisms, and some of it may be done explicitly ("Oops I am being selfish now", "My thinking about this problem is going round in circles", "I don’t really understand what he is saying to me", etc. Part of the process of cultural transmission is giving children the ability to make explicit self-judgements, often harmful ones, as in religious indoctrination.)

Yet more forms of representation/encoding may be involved in other parts of the system. For instance, some sub-mechanisms will use spatial structures as part of the process of perception or problem solving or motor control. These will not map in any simple way onto verbal descriptions of what the animal believes, or thinks, etc. (Likewise some beliefs stored explicitly using the ontology of French may not map well into English, if, as John Lyons has argued, different natural languages can use different sub-ontologies, e.g. for types of household furniture.)

A further problem is that individuals can develop their own ontologies which need not be expressible in the languages used by others.

In summary: if deliberative mechanisms require information to be explicitly chunked in such a way as to support prediction, explanation, and planning, then there may be more truth in the Fodor/GOFAI assumptions than you and connectionists allow, even if there are other parts of the architecture for which those assumptions are completely mistaken.

I believe you would eventually reach this conclusion if you continued developing some of the architectural designs you have already presented, e.g. in Kinds of Minds.

19. (You will agree with most of this?). There is still much philosophical work to be done regarding virtual machines, supervenience, and multiple-realisability.

E.g. the extent of multiple realisability is partly an empirical question, and partly conceptual. The topic is still very muddy.

It is an empirical question how much similarity to human minds could be achieved by a totally different physical implementation, but a partly fuzzy empirical question.

Any particular virtual machine architecture can be described at different levels of detail. The more detail there is the more constraints there will be on possible physical implementations.

There are some people who claim that we shall not be justified in attributing mentality to a working system unless there is not only an appropriate information processing virtual machine, but also an appropriate lower level (or bottom level) implementation machine.

E.g. Steve Grand, creator of the Creatures System, claims that for a system to have mental states and processes there has to be at least a level mirroring the properties of chemical processes of biological systems. Margaret Boden (1999) has claimed that for something to be an instance of "life" there must be some sort of metabolism involved in the implementation. Some people would extend this criterion to mental concepts too, and claim that an implementation involving metabolism is a necessary condition for mentality in a virtual machine.

However, instead of arguing endlessly over which conditions are necessary for life, or for intelligence, or for semantic capabilities, I suggest that all we need to do is analyse the different cases, and draw out the implications of their similarities and differences.

Sometimes the implementation affects important features of the virtual machine. For instance changing computer technology has led to faster, more compact, less energy consuming, more reliable, implementations of many virtual machines.

Sometimes features of the low level implementation machine interact closely with features of a virtual machine, e.g. if the physical mechanisms are particularly unreliable.

The Pop-11 system has a library procedure which finds out the largest size of machine integer the machine it is running on can hold in a machine word. It does this by repeatedly doubling the value of a variable until it no longer satisfies the test for a "small" integer. I.e. the VM detects a limitation which depends on its current physical implementation. (The result is different on 32 bit and 64 bit CPUs.)

A virtual machine can also use sub-mechanisms which detect and react to changes at the implementation level, e.g. if the VM uses information produced by temperature detectors, or mechanisms for detecting physical memory errors which allow the VM to take corrective action (e.g. no longer using a particular portion of physical memory after it has been found to be damaged).

In more complex machines some of the detectors might check whether some energy source or some chemical required for normal functioning is running low, or whether some connections have become unreliable and need to be repaired.

In the extreme case, a high level virtual machine can have components that are closely coupled with very large numbers of physical state monitors and homeostatic mechanisms distributed all over a physical body, some of which are able to interrupt and redirect processing within the virtual machine when certain thresholds are exceeded. If these detectors are mapped onto a model in the virtual machine of the whole body, then such a system may have the experience of being embodied in a particular way. (It will have a rich supply of bodily qualia based on proprioceptive mechanisms.)

The variety of mental states that can occur in such a system will be very different from those in a virtual machine whose architecture is largely disconnected from its underlying physical implementation (e.g. a typical chess playing virtual machine, or a theorem prover).

We may discover that laws of physics constrain the variety of implementations of the former, "richly embodied", sort of system far more than the sort where the virtual machine takes no account of physical states.

In *principle* it may be possible to simulate all the levels of such a richly embodied system within a virtual machine simulating a large portion of the physical universe. However we may discover that the laws of physics rule this out except in very simple cases.

It might then be a contingent, empirical discovery, not a conceptual truth, that certain kinds of mentality in this world can only be implemented in bodies similar to ours, including similar metabolic processes. (I see no reason to believe that physics does impose such a constraint: I merely point out the logical possibility).

If physical reality does have such a restriction, then it would be a fact of physics that only brains like ours could physically support mental states very similar to ours. Different sorts of brains or computers, however, could support different classes of mental states, including some mental states only partly like ours. (Just as men are partly like women, children partly like adults, physicists partly like poets, etc.)

I suspect that most people who argue that only biological brains can support minds are sliding from "physically can" to a broader sense of "can".

Which features characterise human beings is obviously an empirical question, and people differ in their inclinations to use mental concepts to describe various non-human animals and machines with different features from humans. Arguing over which inclinations are *correct* is pointless.

(However one can disagree with the ethics of someone who says that dogs or robots have no pains.)

There are many more details to be sorted out regarding relations between VMs and physical machines. There are many remaining confusions.

For instance, the criteria for individuating physical and virtual machines are indeterminate in different ways. Assuming that they are precisely defined can easily lead to paradoxes (e.g. Merricks, 1998)

E.g. a virtual machine implemented in a computer usually does not require all the parts of that computer to be present. (It may use only some of the available memory, only some of the available physical transducers, and possibly none of the physical casing of the machine.)

There is no well defined subset of the physical machine that has to be present for it to be *that* machine.

Likewise there is no well defined boundary to the set of physical components that constitutes the implementation of the VM currently running on the machine.

Thus, we should not expect the physical body containing a particular mind to be very precisely determined, And conversely, precisely which mind is currently implemented in a particular body will be partly indeterminate.

CONCLUSION

Taking up the design stance as understood by software engineers reveals semantic and other riches not evident to a mere mechanical or physical designer.

It also shows that various questions about what is going on in certain virtual machines are factual questions (though philosophical analysis and conceptual clarification may be required to identify those factual questions where cluster concepts are involved).

There is much conceptual analysis still to be done, e.g. concerning causal relations between VM events and physical events, the requirements for different sorts of implementation (supervenience), and the many large and small discontinuities in domains of application of many of our cluster concepts.

There are also many empirical questions to be investigated about types of designs and implementations that can be supported in the physical world, and the kinds of designs and implementations that are capable of matching typical human minds in more or less detail.

In this context the intentional stance is rarely needed. It is generally more fruitful and provides richer explanatory power to take up the design stance, without any presumption of rationality, even if that involves adopting conjectures that are not supported by the evidence.

But as successful biological organisms, and as good scientists, we are bound to do that anyway.

I could go on, but that's enough for now.

I have the impression from things of yours which I have read over the years that you agree with a great deal of what I've written above, but there are some major disagreements. What I don't know is where the gap starts between us.

My hunch is that you have not yet put together various things you know, and that when you write some of your papers (e.g. the chapter on Real patterns in Brainchildren) you focus on certain phenomena (e.g. we can observe the behaviour of a machine and try to infer what's going on) and when you write other things (e.g. about designs for various types of minds) you focus on what you know about software engineering.

If you put all that together you might change some details of the Real Patterns paper. That's because when you know how something works as a result of being the engineer who designed it and produced the implementation, then you can answer questions which cannot be answered simply on the basis of observations of the system's behaviour, no matter how long you observe it.

Or have I missed something you know about the logic of virtual machines?

NOTE: I'll put a copy of this in

http://www.cs.bham.ac.uk/~axs/misc/draft/dennettletter.html

and invite comments.

If you ever have time to respond, let me know whether you would like me to add your responses to the same directory, with a link from the original.

Cheers.
Aaron
===
[REFS]

M.A.Boden, Is metabolism necessary?, in Brit. Journal for the Philosophy of Science, 50, 2, 1999

D.C. Dennett, Darwin's Dangerous Idea: Evolution and the Meanings of Life, 1995, Penguin Press, London and New York,

D.C. Dennett, Kinds of minds: towards an understanding of consciousness, Weidenfeld and Nicholson, London, 1996,

D.C. Dennett, Brainchildren: Essays on Designing Minds. Penguin Books, 1998.

D. Marr, Vision, Freeman, 1982

T. Merricks, Against the doctrine of microphysical supervenience, in Mind, vol 107, no 425, 1998.

A. Sloman, Why philosophers should be designers, in Behavioral and Brain Sciences, Vol 11 no 3, pp529-530, 1988 (Commentary on D.C. Dennett: The Intentional Stance)

A. Sloman, On designing a visual system (Towards a Gibsonian computational model of vision), in Journal of Experimental and Theoretical AI, 1989, Vol 1, no4, pp. 289--337

A. Sloman, Architectural Requirements for Human-like Agents Both Natural and Artificial. (What sorts of machines can love?), in Human Cognition And Social Agent Technology, Ed. Kerstin Dautenhahn, Advances in Consciousness Research, John Benjamins, Amsterdam, pp. 163--195, 2000

A.Sloman, How many separately evolved emotional beasties live within us? To appear in, Emotions in Humans and Artifacts, Eds Robert Trappl and Paolo Petta, MIT Press. Available now at http://www.cs.bham.ac.uk/research/cogaff/

P. F. Strawson, Individuals: An essay in descriptive metaphysics, Methuen, 1959, London

---

Last updated: 31 May 2000