Machines in the Ghost

Invited Position Paper for ENF 2007 Emulating the Mind
1st international Engineering and Neuro-Psychoanalysis Forum
Vienna July 2007
http://www.indin2007.org/enf/
DRAFT: August 30, 2007(MAY CHANGE)

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

Index Terms— **architecture, artificial-intelligence, autonomy, behaviour, design-based, emotion, evolution, ghost in machine, information-processing, language, machine, mind, robot, philosophy, psychotherapy, virtual machine.**

Abstract— **This paper summarises ideas I have been working on over the last 35 years or so, about relations between the study of natural minds and the design of artificial minds, and the requirements for both sorts of minds. The key idea is that natural minds are information-processing *virtual* machines produced by evolution. What sort of information-processing machine a human mind is requires much detailed investigation of the many kinds of things minds can do. At present, it is not clear whether producing artificial minds with similar powers will require new kinds of computing machinery or merely much faster and bigger computers than we have now. Some things once thought hard to implement in artificial minds, such as affective states and processes, including emotions, can be construed as aspects of the *control* mechanisms of minds. This view of mind is largely compatible in principle with psychoanalytic theory, though some details are very different. The therapeutic aspect of psychoanalysis is analogous to run-time debugging of a virtual machine. In order to do psychotherapy well we need to understand the architecture of the machine well enough to know what sorts of bugs can develop and which ones can be removed, or have their impact reduced, and how. Otherwise treatment will be a hit-and-miss affair.**

I. The design-based approach to studying minds

MY PRIMARY goal is to understand natural minds of all kinds, not to make smart machines. I am more a philosopher than an engineer – though I have first-hand experience of software engineering. Deep scientific and philosophical understanding of natural minds requires us to describe minds with sufficient precision to enable our theories to be the basis for designs for working artificial minds like ours. Such designs can be compared with psychoanalytic and other theories about how minds work. If a theory about how natural minds work is to be taken seriously, it should be capable of providing the basis for the design of artifical working minds.

Part of the problem is describing what needs to be explained, and deciding which concepts to use in formulating explanatory theories. A very serious impediment to progress is the common assumption that we already know what needs to be explained and modelled, leaving only the problem of finding good theories and designs for working systems. Alas, what needs to be explained is itself a topic still requiring much research, as acknowledged by the Research Roadmap project in the euCognition network (http://www.eucognition.org).[1].

A less obvious, but even more serious impediment to progress is the common assumption that our ordinary language is sufficient for describing everything that needs to be explained, leading to over-reliance on common-sense concepts. When scientists and engineers discuss what needs to be explained, or modelled, and when they report experimental observations, or propose explanatory theories, they often use concepts (such as 'conscious', 'unconscious', 'experience', 'emotion', 'learn', 'motive', 'memory') that evolved not for the purposes of science, but for use in informal discourse among people engaged in every day social interaction, like this:

- What does the infant/child/adult/chimp/crow (etc) perceive/understand/learn/intend (etc)?
- What is he/she/it conscious of?
- What does he/she/it experience/enjoy/desire?
- What is he/she/it attending to?
- What sort of emotion is he/she/it having now?

These everyday usages may be fine for everyday chats, gossip, consulting rooms, or even law courts and medical reports, but it does not follow that the concepts used (e.g. 'conscious', 'experience', 'desire', 'attend', 'emotion') are any more adequate as concepts to be used in scientific theories than the everyday concepts of 'mud', 'cloudy', 'hot', 'vegetable', 'reddish' or 'smelly' are useful in theories about physics or chemistry.

From vernacular to deep concepts

The history of the physical sciences shows that as we understand more about the architecture of matter and the variety of states, processes, and causal interactions that lie hidden from ordinary observation, the more we have to construct new systems of concepts and theories that make use

This work was partly supported by the EU CoSy project.

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

[1] Available online: 'What's a Research Roadmap For? Why do we need one? How can we produce one?' http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0701

of them (and thereby help to define the concepts), in order to obtain deep explanations of what we can observe and new more reliable and more precise predictions, regarding a wider range of phenomena. In particular, in physics, chemistry and biology, new theories have taught us that however useful our ordinary concepts are for ordinary everyday social interactions, they are often *grossly* inadequate: (a) for distinguishing all the phenomena that need to be distinguished, (b) for formulating precise and reliable predictions, and (c) for describing conditions (often unobservable conditions) that are the basis for reliable predictions.

So it may be unwise to go on using the old concepts of folk psychology when describing laboratory experiments or field observations, when formulating descriptions of what needs to be explained, and when formulating supposedly explanatory theories. But choosing alternatives to ordinary language needs great care. Recently attempts have been made to give these concepts scientific status by using modern technology to identify precise brain mechanisms, brain states, brain processes that correlate with the states and processes described in ordinary language. Compare trying to identify the precise location in a particular country where some national characteristic or process such as religious bigotry, scientific ignorance, or economic inflation is located.

The 'design-based' approach

Is there any alternative to going on using pre-scientific contexts in our descriptions and theories? Yes, but it is not a *simple* alternative. We need to adopt what Dennett [1] calls 'the design stance', which involves constructing theories about how minds and brains actually work, which goes far beyond what we either experience of their working in ourselves or observe in others. (My own early attempts at doing this 30 years ago are presented in [2].) Moreover, our theories must account for the existence of many kinds of minds, with different designs.

In particular we can't just start defining precise new explanatory concepts in terms of precise measurements and observations. Thinking that definitions come before theories is a common mistake. In the more advanced sciences, concepts are used (e.g. 'electron', 'charge', 'atomic weight', 'valency', 'oxidation', 'gene', etc.) that cannot be defined except by their role in the theories that employ them. The *structural relations* within the theory partially define the concepts. That is because a theory is a formal system, and, as is familiar from mathematics, and made more precise in the work of Tarski, the structure of a formal system determines which things are possible *models* of that system. Usually there are many possible models, but the set of possible models can be reduced by enriching the theory, thereby adding more constraints to be satisfied by any model. This may still leave different models, of which only one, or a subset is intended. Such residual ambiguities are reduced (but never completely removed) by links between the theory and methods of observation and experiment, and practical applications associated with the theory. However those links do not *define* the concepts, because old methods of observation and old experimental procedures can be replaced by new ones, while the old concepts endure. This loose, theory-mediated, connection between theoretical concepts and observable phenomena is referred to as 'symbol attachment' in [3] and 'symbol tethering' in [4]. Once this possibility is understood, the need for so-called 'symbol grounding' (deriving all concepts from experience) presented in [5], evaporates. (The impossibility of 'concept empiricism' was demonstrated by Kant [6] over 200 years ago.)

The ontology used by an individual or a community (i.e. the set of concepts used to describe things in the world) can be extended in two ways, either by definitional abbreviation or substantively. A definitional abbreviation merely introduces a new symbol as a short-hand for what could be expressed previously, whereas *substantive* ontology extensions introduce new concepts that cannot be defined in terms of pre-existing concepts. Such concepts are implicitly defined mainly by their role in explanatory theories, as explained above. So *substantive* ontology extension always requires theory construction, as has happened many times in the history of science and culture.

II. Virtual and physical machines

We can do for the study of mind and brain what was previously done for the study of physical matter and biological processes of evolution and development, if we understand that minds and brains are not just matter-manipulating, or energy-manipulating machines, but information-processing machines. Minds are information processing *virtual machines* while brains are *physical* machines, in which those virtual machines are *implemented* or some would say *realised*.

In computing systems we also have virtual machines, such as running operating systems, firewalls, email systems, spelling correctors, conflict resolution mechanisms, file optimisation mechanisms, all of which run on, i.e. are implemented in, the underlying physical hardware. There are no simple mappings between the components of the virtual and the physical machines. There are often several layers of implementation between high level virtual machines and physical machines. This provides enormous flexibility for re-use of hardware, both in different systems using the same hardware, and in one system performing different functions at different times. It seems that evolution discovered the power of virtual machines before we did and produced brains implementing hierarchies of mental (and indirectly social) virtual machines.

The various components of a complex virtual machine (such as an operating system distributed over a network of processors) need not run in synchrony. As a result, timing relations between processes can change from time to time. For this and other reasons, Turing machines do not provide good models for minds, as explained in [7]. Moreover if sophisticated memory management systems are used the mappings between components of virtual machines and physical parts of the system can also keep changing.

If such complex changing relationships are useful in systems we have designed and implemented we need to keep an open mind as to whether evolution, which had several billion years head start on us, also discovered the power and usefulness of

such flexibility. If so, some apparently important searches for mind-brain correlations may turn out to be a waste of time.

Can virtual machines do things?

It is sometimes thought that the virtual machines in a computer do nothing: they are just figments of the imagination of software engineers and computer scientists. On this view, only the physicists and electronic engineers really know what exists and interacts in the machine. That derives from a widely held theory of causality, which assumes either that only physical events can really be causes that produce effects, or, more subtly that if events are caused physically then they cannot be caused by events in virtual machines. Some researchers have inferred that human mental events and processes, such as weighing up alternatives, and taking decisions cannot have any consequences: they are mere epiphenomena. There is no space here for a full rebuttal of this view, but the key idea is that causation is not like some kind of fluid or physical force that flows from one thing to another. Rather, the concept of X causing Y is a very subtle and complex concept that needs to be analysed in terms of whether Y would or would not have happened in various conditions if X did or did not occur, or if X had or had not occurred, or if X does or does not occur in the future. The truth or falsity of those 'counterfactual conditional' statements depends in complex ways on the truth or falsity of various laws of nature, which we attempt to express in our explanatory theories.

However, in addition to their role in true and false statements counterfactual conditionals may also play a role in instructions, intentions, or motives. Thus 'Had he not cooperated I would have gone ahead anyway' may be not so much a retrospective *prediction* as an *expression of resolve*. Plans, intentions and strategies involve causation as much as predictions and explanations do. These too can play a causal role in virtual machines.

Real causation in virtual machines

Suppose one of your files disappears. A software engineer may conclude (after thorough investigation) that one of your actions activated a bug in a running program, which led another part of the program to remove that file. These are events and processes in a virtual machine. This diagnosis could lead the programmer to make a change in the operating system or file management system that alters its future behaviour. Sometimes this can be done by altering software rules without restarting the machine – rather like telling a person how to do better.

Typically, the causal connections discovered and altered by software engineers do not require any action by electronic engineers or physicists to alter the physical machine, and most of the people with expert knowledge about the hardware would not even understand how the bug had occurred and how it was fixed.

Moreover, virtual machine components can constantly change their mapping onto hardware, so the change made by the software engineer need not have any specific physical location when the software is running.

The mind-brain identity defence

Some philosophers defend the thesis that 'only *physical* events can be causes' by claiming that the objects, events and processes in virtual machines actually *are* physical objects, events and processes viewed in an abstract way. This is a 'Mind-brain identity theory', or a 'virtual-physical machine identity theory', sometimes presented as a 'dual-aspect' theory. A detailed rebuttal of this thesis requires showing (a) that it is based on a bad theory of causation, (b) that the claim of identity being used here is either vacuous or false because the actual relations between contents of virtual machines and contents of physical machines are quite unlike other cases where we talk of identity (e.g. because the virtual to physical mapping constantly changes even within the same machine) (c) that the concepts used in describing virtual machine phenomena (e.g. 'software bug', 'failure to notice', 'preferring X to Y', etc.) are too different from those of the physical sciences to be capable of referring to the same things. More obviously, those virtual machine concepts are not *definable* using only concepts of the physical sciences.

This is a very compressed defence of the claim that virtual machine events can have causal powers. There are more detailed discussions on the Birmingham Cogaff web site.[2]

Psychotherapy as virtual machine debugging

This design-based approach using the concept of a 'virtual machine' can, in principle, justify techniques that deal with a subset[3] of human mental problems by manipulating virtual machines instead of manipulating brains using chemicals, electric shocks, etc. If a virtual machine is suitably designed then it is possible to identify and in some cases repair, certain 'bugs', or 'dysfunctional' processes, by interacting with the running system. Such debugging is common in teaching.

Doing that well requires *deep, explanatory* theories of how normal mental virtual machines work. Some debugging techniques for minds have evolved through various kinds of social experimentation without such deep theories, and many of them work well. For instance, when a child gets the wrong results for long division we don't call in a brain surgeon, but check whether he has perhaps learnt the wrong rule or misunderstood one of the mathematical concepts involved. In such cases the lack of competence can be seen as a bug that can be fixed by talking, drawing diagrams, and giving simpler examples. It might be thought that this is unlike fixing bugs on computers because in the latter case the process has to be stopped, the new program compiled and the program restarted whereas humans learn and change without having to cease functioning. But some programs are run using interpreters or incremental compilers, which allow changes to be made to the running program *without* stopping and restarting the process with a recompiled program. Several AI programming languages have that kind of flexibility. More sophisticated software development tools can also plant mechanisms for interrogating and modifying running systems.

Of course, teaching someone how to improve his long

[2] See this presentation on virtual machines `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#inf` and this discussion of free-will and causation `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/four-kinds-freewill.html`

[3] Don't expect it to remove brain-tumours, for instance!

division, and giving counselling to enable a patient to understand how he unintentionally causes family rows to escalate, differ in detail from the process of fixing a typical bug in a computer program. For example the teaching and therapy depend on kinds of self understanding that few computer programs have at present. But in future it will not be uncommon to find virtual machines, that, instead of being forcibly altered by a user editing rules, instead change themselves as a result of being given *advice* about how to behave. That's not even science fiction: it is not hard to achieve. This comparison should undermine three assumptions: (a) the belief that the only way to understand and fix problems with minds is to work in a totally bottom up way, namely understanding and modifying brains, (b) the belief that nothing we know about computers is relevant to understanding and repairing minds, and (c) the belief that machines can only be *programmed* to do things, not advised, inspired, or instructed or cajoled into doing them.

Motivation in virtual machines

So, states, events, and processes occurring in virtual machines can have causal influences which alter both other things in the virtual machine and also the physical behaviour of the system (e.g. physical events in memory, hard drives, internal interfaces, what is displayed on a screen, the sounds coming from speakers, or various attached motors, etc.). However, we still need to understand the variety of kinds of causation in virtual machines.

In very simple computer models nothing happens until a user gives a command, which can then trigger a cascade of processes. In more sophisticated cases the machine is designed so that it initiates various activities from time to time, e.g. checking for email, checking whether disks need defragmenting, checking whether current scheduling parameters need to be revised in order to improve processing performance. It can also be designed so that events initiated from outside trigger new internal processes. E.g. a user attempting to access a file can trigger a sub-process checking whether the user has the right to access that file. All of these cases require the designers of the virtual machine to anticipate kinds of things that might need appropriate checking or corrective actions to be performed.

But there is no difficulty in building a machine that acquires new competences while it is running, and also new conditions for exercising old competences. If the virtual machine architecture allows new goal-generators to be acquired, new strategies for evaluating and comparing goals, new values to be employed in such processes, then after some time, the machine may have goals, preferences, intentions, etc. that were not given to it by anyone else, and which, as remarked in [2], can only be described as its own. It would be, to that extent, an *autonomous* machine, even if the processes by which such motives and motive-generators were acquired involved being influenced by things said and done by other people, for instance teachers, heroic figures, and other role models. Indeed a machine might be designed specifically to derive motives, values, preferences, etc. partly on the basis of such influences, tempered by experimentation on the results of trying out such values. Isn't that what happens to humans? This point was made long ago in section 10.13 of [2]. So motivation in artificial virtual machines, including self-generated motivation, is not a problem in principle.

The myth that intelligence requires emotions

In recent years, especially following publication of Damasio's [8] and Picard's [9], much has been made of the alleged need to ensure that intelligent systems have emotions. I have argued elsewhere that the arguments are fallacious for example in [10] [11] [12] [13] [14] and [15]. Some of these claims are merely poorly expressed versions of Hume's unsurprising observation that without *motives* an intelligent system will have no reason to *do* anything: this is just a confusion between emotions and motives.

There is also a more subtle error. 26 years ago, [10] argued that intelligent machines need mechanisms of kinds that perform important functions and which *in addition* can sometimes generate emotional states and processes as *side effects* of their operation, if they lack sufficient processing power to work out what needs to be done. This does not imply that they *need* emotions,[4]. just as the fact that some operating systems need mechanisms that are sometimes capable of generating 'thrashing' behaviour does not imply that operating systems *need* thrashing behaviour. Desirable mechanisms can sometimes have undesirable effects. There can be particular sorts of situations where an 'alarm system' detects a putative need to override, modulate, freeze, or abort some other process, and where because of shortage of information or shortage of processing power, a rough and ready rule operates. It would be better if the situation could be fully evaluated and reasoned about (without any emotions), but if there is inadequate capacity to do that in the time available then it may be better to take the risk of false alarms, especially if the alarm system has been well trained, either by evolution or individual learning, and does not often get things wrong.

The need for therapy to undo bad learning

However the danger in having such powerful subsystems that can change as a result of learning is that they may change in bad ways – as clearly happens in some humans. In some cases, it may be possible to undo bad changes by re-programming, e.g. through discussion, advice, teaching, re-training, therapy, etc. In other cases help may be available only when it is too late. Of course there are also cases where such systems go wrong because of physical damage, disease, malfunction, corruption, etc., which may or may not be reversible, e.g. chemical addictions.

III. The variety of mental virtual machines

There is not just one kind of mind. Insect minds are different from minds of birds and monkeys. All are different in many ways from adult human minds. Human minds are different at different stages of development: a newborn infant, a nine-month old crawler, a two-year old toddler, a four year old talker, a 50 year old professor of psychiatry. Apart from

[4]As argued in this slide presentation: `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cafe04`

differences resulting from development and learning there are differences that can be caused by genetic deficiencies, and by brain malfunctions caused by disease or injury. What's more obvious is that the brains are different too. So we need to find a general way of talking about different minds, different brains, and the different kinds of relationships that hold between (a) the minds, i.e. the virtual machines, and (b) the brains, i.e. the physical machines in which they are implemented.

Even virtual machines have architectures

We are talking about a type of complex system with many concurrently active parts that work together more or less harmoniously most of the time but can sometimes 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 we should ask questions like this if we wish to do deep science studying a particular kind of mind:

- What sorts of component parts make up the architecture of this sort of mind?
- What are their functions?
- For each such component, what difference would it make if it were modified or removed, or connected in a different way to other components?
- Which parts of the architecture are involved in various processes that are found to occur in the system as a whole?
- Which parts are connected to which others, and how do they interact?
- What kinds of information do the different parts acquire and use, and how do they obtain the information?
- How is the information represented? (It could be represented differently in different subsystems).
- Can the system extend the varieties of information contents that it can make use of (extend its ontology), and extend the forms of representation that it uses?
- What is the total architecture in which they function, and how is it made up of sub-architectures?
- How are the internal and external behaviours selected/controlled/modulated/coordinated?
- Can conflicts between subsystems arise, and if so how can they be detected, and how can they be resolved?
- What mechanisms in virtual machines make those processes possible, and how are they implemented in brains?
- In how many ways can the different virtual machines either individually, or through their interactions go wrong, or produce dysfunctional effects?
- How did this virtual machine evolve, and what does it have in common with evolutionary precursors and with other contemporary animal species?

Answering these, and similar, questions requires a long term investigation. One of the reasons why optimistic predictions regarding imminent successes of Artificial Intelligence (or more recently robotics) have repeatedly failed is that the phenomena we are trying to explain and to replicate, are far more complex than anyone imagines. I.e. the main reason is NOT that the wrong programming languages, or the wrong models of computation were used (e.g. symbolic *vs* neural), or that the test implementations used simulations rather than real robots, but that what the researchers were trying to get their systems to do fell very far short of what they implied in their predictions and promises would be achieved, because they had not analysed the requirements adequately.

Moreover, it could turn out that *all* of the currently understood models of computation are inadequate, and entirely new kinds of virtual machine are needed, including machines that grow their own architecture, as suggested in [3].

IV. What is information?

There are many questions still to be answered about the concept of information used here. What is 'information'? What is an information-user? What is involved in understanding something as expressing a meaning or referring to something? Is there a distinction between things that merely manipulate symbolic structures and things that manipulate them while understanding them and while using the manipulation to derive new information? In how many different ways do organisms acquire, store, derive, combine, manipulate, transform and use information? How many of these are, or could be, replicated in non-biological machines? Is 'information' as important a concept for science as 'matter' and 'energy', or is it just a term that is bandied about by undisciplined thinkers and popularists? Can it be defined? Is information something that should be measurable as energy and mass are, or is it more like structure, which needs to be described not measured (e.g. the structure of this sentence, the structure of a molecule, the structure of an organism, the properties of a toroid)?

We currently have the paradoxical situation that philosophers who are good at conceptual analysis are badly informed about and often have false prejudices about computing systems, whereas many software engineers have a deep but unarticulated understanding, which they use very well when designing, implementing, testing, debugging, maintaining, virtual machines, but which they cannot articulate well because they have not been taught to do philosophy.

Information-theory is not about information!

The word 'information' as used (after Shannon) in so-called 'information theory' does not refer to what is normally meant by 'information', since Shannon's information is a purely syntactic property of something like a bit-string, or other structure that might be transmitted from a sender to a receiver using a mechanism with a fixed repertoire of possible messages. Having that sort of information does not, for example, allow something to be true or false, or to contradict something else. However, the more general concept of information, like 'mass', 'energy' and other deep concepts used in scientific theories, is not explicitly definable. That is to say, there is no informative way of writing down an explicit definition of the form 'X is Y' if X is such a concept. All you'll end up with is something circular, or potentially circular, when you expand it, e.g. 'information is meaning', 'information is semantic content', 'information is aboutness', 'information is what is expressed by something that refers', 'information is a difference that makes a difference' (Bateson), and so on.

But that does not mean either that the word is meaningless or that we cannot say anything useful about it. The same is true

of ‘energy’. It is sometimes defined in terms of ‘work’, but that eventually leads in circles. So how do we (and physicists) manage to understand the word ‘energy’?

The answer was given above in Section I: we understand the word ‘energy’ and related concepts, by understanding their role in a rich, deep, widely applicable theory (or collection of theories) in which many things can be said about energy, e.g. that in any bounded portion of the universe there is a scalar (one-dimensional), discontinuously variable amount of it, that its totality is conserved, that it can be transmitted in various ways, that it can be stored in various forms, that it can be dissipated, that it flows from objects of higher to objects of lower temperatures, that it can be used to produce forces that cause things to move or change their shape, etc. etc. As science progresses and we learn more about energy the concept becomes deeper and more complex. The same is happening with ‘information.’

An implicitly defined notion of ‘information’

We understand the word ‘information’ insofar as we use it in a rich, deep, and widely applicable theory (or collection of theories) in which many things are said about information, e.g. that it is not conserved (I can give you information without losing any), that instead of having a scalar measure of quantity, items of information, may form a partial ordering of containment (information I2 is contained in I1 if I2 is derivable from I1), and can have a structure (e.g. there are replaceable parts of an item of information such that if those parts are replaced the information changes but not necessarily the structure), that two information items can share some parts (e.g. ‘Fred hates Mary’ and ‘Mary hates Joe’), that it can be transmitted by various means from one location or object to another, that it can vary both discontinuously (e.g. adding an adjective or a parenthetical phrase to a sentence, like this) or continuously (e.g. visually obtained information about a moving physical object), that it can be stored in various forms, that it can influence processes of reasoning and decision making, that it can be extracted from other information, that it can be combined with other information to form new information, that it can be expressed in different syntactic forms, that it can be more or less precise, that it can express a question, an instruction, a putative matter of fact, and in the latter case it can be true or false, known by X, unknown by Y, while Z is uncertain about it, etc. etc.

Information is relative to a user, or potential user

Whereas energy and physical structures simply exist, whether used or not, information in a physical or virtual structure S is only information for a type of user. Thus a structure S refers to X or contains information about X *for a user of S, U*. The very same physical structure can contain different information for another user U’, or refer to something different for U’, as shown by ambiguous figures, and also written or spoken languages or notations that some people understand and others do not. This does not make information inherently subjective any more than mountains are subjective because different people are capable of climbing different mountains. (Indexicality is a special case, discussed below.) The information in S can be potentially usable by U even though U has never encountered S or anything with similar information content, for instance when U encounters a new sentence, diagram or picture for the first time. Even before any user encounters S, it is *potentially* usable as an information bearer. Often, however, the potential cannot be realised without U first learning a new language, or notation, or a new theory within which the information has a place, and which provides substantive ontology extension for U, as discussed in [16]. This may be required for growth in self-knowledge too.

A user with appropriate mechanisms has potential to derive infinitely many distinct items of information from a small structure, e.g. infinitely many theorems derivable from Peano’s five axioms for arithmetic. Physically quite small objects can therefore have infinite information content, in combination with a reasoning mechanism, though limitations of the implementation (e.g. amount of memory available) may constrain what is actually derivable. It follows that physical structure does not constrain information content, unless a type of user is specified.

Information processing in virtual machines

Because possible operations on information are much more complex and far more varied than operations on matter and energy, engineers discovered, as evolution had ‘discovered’ much earlier, that relatively unfettered information processing requires use of a virtual machine rather than a physical machine. E.g. digital electronic calculators can perform far more varied tasks than mechanical calculators using cog-wheels.

It seems to be a basic law that increasing usable information in a virtual machine by making implications explicit requires the physical implementation machine to use energy. Similarly (as suggested by Jackie Chappell), using greater information content requires more energy to be used: e.g. in storage, sorting and processing information. So biological species able to acquire and process vast amounts of information must be near the peak of a food pyramid, and therefore rare.

Causal and correlational theories of meaning are false

It is often thought that learning to understand S as referring to X, requires an empirical discovery that there is a causal relation between S and X, such as that occurrences of X always or often cause occurrences of S to come into existence, or such that the occurrence of X is shown empirically to be a reliable predictor of the occurrence of S. That this theory is false is shown by the fact that you have no reason to believe that occurrences of the word ‘eruption’ are correlated with or reliable predictors of eruptions, and moreover you can understand the phrase ‘eruption that destroyed the earth 3000 years ago’ even though it is impossible for any such correlation or causal link to exist, since the earth was not destroyed then. Further we can use concepts that refer to abstract entities whose existence is timeless, such as the number 99, or the shortest proof that 2 has no rational square root. So empirical correlations and causal influences are impossible in those cases.

Information content determined partly by context

It is sometimes thought that artificial minds would never

be able to grasp context-sensitive information. For example, an information-bearing structure S can express different information, X, X', X", for the same user U in different contexts, e.g. because S includes an explicit *indexical* element (e.g. 'this', 'here', 'you', 'now', or non-local variables in a computer program). Indexicality can make information incommunicable in the sense that the precise information content of one user cannot be transferred to another. (Frege, for example, showed that one user's use of the word "I" has a sense that another person is incapable of expressing.) A corollary is that information acquired by U at one time may not be fully interpretable by U at another time, because the context has changed, e.g. childhood 'memories'. In [17] it was argued that such indexicality accounts for the 'ineffability' of qualia. However this does not usually prevent the *intended function* of communication from being achieved. The goal of communication is not to replicate the sender's mental state, or information content, in the receiver, but to give the receiver information that is adequate for some purpose.

Many structures in perceptual systems change what information they represent as the context changes. Even if what is on your retina is unchanged after you rapidly turn your head 90 degrees in a room, the visual information will be taken to be about a different wall – with the same wallpaper as the first wall. Many examples can be found in [18].

Sometimes U takes S to express different meanings in different contexts because S includes a component whose semantic role is to express a higher order function which generates semantic content from the context, e.g. 'He ran after the smallest pony'. Which pony is the smallest pony can change as new ponies arrive or depart. More subtly what counts as a tall, big, heavy, or thin something or other can vary according to the range of heights, sizes, weights, thicknesses of examples in the current environment.

There are many more examples in natural language that lead to incorrect diagnosis of words as vague or ambiguous, when they actually express precise higher order functions, applied to sometimes implicit arguments, e.g. 'big', 'tall', 'efficient', 'heap', or 'better' (discussed in [19]).[5]

Information content shared between users

Despite the above, it is sometimes possible for X to mean the same thing to different users U and U', and it is also possible for two users who never use the same information bearers (e.g. they talk different languages) to acquire and use the same information. This is why relativistic theories of truth are false: although I can believe that my house has burned down while my neighbour does not, one of us must be mistaken: it cannot be true for me that my house has burned down but not true for my neighbour. Truth is not 'for' anyone. Meaning depends on the user. Truth does not.

V. Information-using subsystems

An information-user can have parts that are information users. A part can have and use some information that other parts cannot access. When we ask 'Did X know that P?' it is not clear whether the answer must be 'yes' in cases where some part of X made use of the information that P. E.g. human posture control mechanisms use changes in optical flow. Does that mean that when you are walking around you know about optical flow even though you don't know that you know it? Your immune system and your digestive system and various metabolic processes use information and take decisions of many kinds. Does that mean that *you* have the information, that you know about the information, that you use the information? Some people might say yes others no, and some may say that it depends on whether you know that you are using the information.

Likewise there are different parts of our brains that evolved at different times that use different kinds of information (even information obtained via the same route, e.g. the retina or ear-drum, or haptic feedback). Some of them are evolutionarily old parts, shared with other species (e.g. posture control mechanisms), some newer and possibly some unique to humans, (e.g. human face recognition mechanisms, and mechanisms that can learn to read music and other notations).

A deep feature of at least the human architecture, and perhaps some others also, is that sensors and effectors providing interfaces to the environment can be shared between different subsystems, as described in [20]. E.g. your vision mechanisms can be shared between: posture control, visual servoing of manipulation actions, mechanisms involved in reading instructions, and affective mechanisms that appreciate aesthetic qualities of what is seen. Your walking mechanisms can be shared between a subsystem concerned with moving to the door and also a subsystem concerned with social communication, e.g. flirting by walking suggestively. We can describe such architectures as using 'multi-window' perception and 'multi-window' action, whereas current artificial systems mostly use only 'peephole' perception and 'peephole' action, where input and output streams from each sensor or two each effector are go along channels of restricted functionality.

Sometimes the sharing is concurrent and sometimes sequential. Conflict resolution mechanisms may be required when concurrent sharing is impossible. Some of the mechanisms that detect and resolve conflicts may be inaccessible to self-monitoring (discussed later), so that an individual may be unaware of important decisions being taken.

Much philosophical, psychological, and social theorising misguidedly treats humans as unitary information users, including Dennett's intentional stance and what Newell refers to as 'the Knowledge level'.

Just the beginning of an analysis of 'information'

The analysis of the concept of 'information' presented here amounts to no more than a small fragment of the full theory of types of states, events, processes, functions, mechanisms and architectures that are possible in (virtual and physical) information-processing machines. I doubt that anyone has produced a clear, complete and definitive list of facts about information that constitute an implicit definition of how we (the current scientific community well-educated in mathematics, logic, psychology, neuroscience,

[5]This idea is developed in the context of Grice's theory of communication, with implications for the evolution of language, here: `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0605` Spatial prepositions as higher order functions.

biology, computer science, linguistics, social science, artificial intelligence, physics, cosmology, ...) currently understand the word 'information'.

E.g. there's a great deal still to be said about the molecular information processing involved in development of an individual from a fertilised egg or seed, and in the huge variety of metabolic processes including intrusion detection, damage repair, transport of materials and energy, and control by hormones and neurotransmitters. It may be that we shall one day find that far more of the brain's information processing is chemical than anyone dreams now is possible.

A more complete theory would provide a more complete implicit definition of the concept 'information' required for understanding natural and artificial systems (including far more sophisticated future artificial systems). A hundred years from now the theory may be very much more deep and complex, especially as information processing machines produced by evolution still seem to be orders of magnitude more complex than any that we so far understand.[6]

VI. Biological information processing

All living things, including plants and single-celled organisms, process information insofar as they use sensors to detect states of themselves or the environment and use that information either immediately or after further processing to select from a behavioural repertoire. The behaviour may be externally visible physical behaviour or internal processes.

While using information an organism normally also uses up stored energy (usually chemical energy), so that it also needs to use information to acquire more energy.

There are huge variations both between the kinds of information contents used by different organisms and between different ways in which information is acquired, stored, manipulated and used by organisms. The vast majority of organisms use only two kinds of information: (a) *genetic information* acquired during evolution, used in replication, physical development and maintenance of the individual, and (b) *transiently available* information used by online control systems. I call the latter 'implicit' information. There may also be some less transient implicit information produced by gradually adjusted adaptive mechanisms.

Since the vast majority of species are micro-organisms the vast majority of information-using organisms can use only implicitly represented information; that is to say they use only information that is available during the transient states of activation produced by information being acquired and used, and the information is represented only *transiently* in activation states of sensors, motors and intervening mechanisms, and also in parameters or weights modified by adaptive feedback mechanisms.

The short term transient implicit information in patterns of activation and the longer term implicit information in gradually changing adaptive mechanisms together suffice for most living things – most of which do not have brains! In some animals brains, add only more of the same. However, in humans and many other animals there additional kinds of information and information processing.

Brains are needed for more than movement

Most organisms manage without brains. Why not all? One function is resolving conflicts between different parts responsible for decisions that could be incompatible, e.g. decisions to move one way to get to food or to move another way to avoid a sensed predator. This requires coordination, possibly based on dynamically changing priorities. A different kind of requirement, discussed later, is doing more complex processing of information, e.g. in order to acquire a better understanding of the environment, or to acquire something like a terrain map, or to plan extended sequences of actions. In organisms with many complex parts performing different functions it may also be necessary to coordinate internal changes, for example all the changes involved in reaching puberty in humans, or during pregnancy, where many internal changes and external behaviours need to be coordinated.

Lewis Wolpert wrote, in an *Observer* book review, March 24, 2002: *'First the only function of the brain from an evolutionary point of view is to control movement and so interaction with the environment. That is why plants do not have brains.'*[7] This often quoted, but grossly misleading, claim reflects a widespread current focus on research that models brains as sensory-motor control mechanisms, e.g. [21]. Even if the original function of brains was to control movement, much of what human brains do has nothing to do with control of movement – for example explaining what is observed, predicting future events, answering a question, and doing mathematics or philosophy.

So we see that information in organisms may be implicit and transient, implicit and enduring, explicit and capable of multiple uses, used only locally, used in controlling spatially distributed information processing, and may vary in kind of abstraction, in the ontology used, in the form of representation used and in the kinds of manipulation that are available.

Varieties of explicit information

Only special conditions bring about evolution of mechanisms that create, store and use *explicit*, that is *enduring, re-usable*, information structures that can be used in different ways (e.g. forming generalisations, making predictions, building terrain maps, forming motives, making plans, remembering what happened when and where, communicating to other individuals, etc.) As explained in [22], this requires new architectures involving mechanisms that are not so directly involved in the sensorimotor relationships and their control.[8]

There may have been intermediate stages of evolution in which non-transient information was stored only in the environment, e.g. in chemical trails and land-marks used as sources of information about the presence of food or predators, or about the routes followed by con-specifics (pheromone trails), or about location relative to a nest or an enduring source

[6]For a draft attempt to answer the question 'What is information?' see http://www.cs.bham.ac.uk/research/projects/cogaff/misc/whats-information.html

[7]http://observer.guardian.co.uk/science/story/0,,673268,00.html

[8]See 'Sensorimotor vs objective contingencies' http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0603

of food. (Some insects can use land-marks so this capability probably evolved a long time ago.) This is possible only for animals in an environment with stable structures, unlike some marine environments.

Mechanisms that evolved to use external enduring information may have been precursors to mechanisms using internally stored explicit information. There probably were many different evolutionary transitions, adding extra functionality. In previous papers (e.g. [23]) colleagues and I have emphasised three main categories of competence requiring different sorts of architectural components, namely reactive, deliberative and meta-management capabilities, but there are many intermediate cases and different sorts of combinations of cases that need to be understood – not only for understanding how things work when everything functions normally, but also in order to understand the many ways things can go wrong, including both the consequences of physical malfunctions and also the consequences of dysfunctional processing of information in virtual machines.

Deliberative mechanisms

A very small subset of organisms (and some machines) have a 'deliberative' information-processing capability insofar as they can construct a set of information structures of varying complexity (e.g. plans, predictions, theories), then compare their merits and de-merits in relation to some goal, and produce new information structures describing those tradeoffs, then select one of the alternatives and make use of it. That ability to represent and reason about hypotheticals was one of the first aspects of human processing modelled in AI subsystems, but it was probably one of the last to evolve, and only a very tiny subset of organisms have it in its richest form, described in [24].

It is now fashionable to contrast that early 'symbolic' AI work with *biologically inspired* AI work. But that is just silly, since all of the work was biologically inspired insofar as it was an attempt to model processes that occur in human beings. A human making a plan or proving a theorem is just as much a biological organism as a human running, jumping or catching a ball, even though more species share the capabilities required for the latter activities. Biology is much richer than some researchers seem to realise.

Meta-semantic competences

Another relatively rare biological information-processing capability involves the ability to refer to, reason about, or care about things that themselves contain, or use information. For instance when you discover what someone thinks, and when you worry about someone's motives you are using such meta-semantic competence. Humans can also apply this *meta-semantic* competence to themselves (though probably not at birth: the architecture needs time to develop). Having this ability requires a more complex architecture than merely being able to refer to things without semantic content (e.g. physical objects and processes). That's partly because having meta-semantic content involves representing things that are treated as true in only a hypothetical, encapsulated, way. The very same form of representation may be used for what an individual A believes is the case and for what A thinks another individual B believes is the case, but the functional roles of those two forms of representation will be different.[9] In particular, the former will be 'referentially transparent' and the latter 'referentially opaque', and the conditions for truth of the two beliefs are quite different – something young children do not learn in the first few years.

The ability to think and reason about the mental contents of another has much in common with the ability to think and reason about one's own mental contents as far as mechanisms and formalisms are concerned, though obviously different sensor mechanisms are involved, i.e. external and internal. However the evolutionary benefits are very different. It is not clear which evolved first. Probably both types of meta-semantic competence co-evolved, each helping to enrich the other.

Incidentally, nothing said here implies that any organism has *full* self-knowledge. On the contrary, self-knowledge will inevitably be associated with a bottleneck with limited capacity – a point that has relevance for psycho-analysis. Compare [25].

Perhaps a more widely-shared ability to formulate internal questions or goals was an evolutionary precursor to meta-semantic competence, since formulating a yes-no question ('Will it rain today?') requires an ability to represent propositions that are capable of being true or false without a commitment to their truth value. The same is true of having desires and intentions. So the ability to represent 'X thinks I will eat him' and to reason about the consequences of X's thinking that (e.g. X will run away) may have arisen as a modification of the ability to represent one's own goals ('I will eat X') that have not been achieved, plans that have not yet been carried out, tentative predictions that have not yet been tested, and questions that have not yet been answered.

Old and new in the same architecture

It is not always remembered that besides having such sophisticated (and possibly unique) capabilities for explicit manipulation of information, humans share many information processing capabilities with other species that lack the distinctively human capabilities. For instance many animals have excellent vision, excellent motor control, abilities to cooperate on certain tasks, hunting capabilities, the ability to avoid predators, nest-building capabilities, as well as the information-processing involved in control of bodily functions. The notion that somehow human cognition, or human conscious processes can be studied and replicated without taking any account of how these more general animal mechanisms work and how they interact with the distinctively human mechanisms, may lead both to a failure to understand humans and other animals and also to designs for robots that don't perform as required.

Some of the ideas developed with colleagues in Birmingham about the different components in such a multi-functional architecture are reported in the paper by Brigitte Lorenz and Etienne Barnard presented at this conference and will not be

[9] J. Barnden's ATT-META project has developed a way of making that distinction which is also related to the ability to think metaphorically. `http://www.cs.bham.ac.uk/~jab/ATT-Meta/`

developed here. From this standpoint it is almost always a mistake to ask questions like 'How do humans do X?' Instead we should ask 'How do different subsystems in humans do X?' (e.g. X = control actions, interpret visual input, learn, store information, react to interruptions, generate motives, resolve conflicts, etc.) And we should expect different answers not only for different subsystems, as in Trehub's [26], but also for humans at different stages of development, in different cultures, with and without brain damage, etc.

VII. Pre-linguistic competences

It is often assumed that there is a massive discontinuity between human linguistic competence and other competences (e.g. as argued by Chomsky in [27]), though there are many who hotly dispute this (e.g. Jablonka and Lamb [28]). But people on both sides of the dispute make assumptions about human language that may need to be challenged if we are to understand what human minds are, and if we wish to produce working human-like artificial minds. In particular it is often assumed that the essential function of language is communication. But, as argued in [29], and more recently in [30] many of the features that make such communication possible (e.g. the ability to use varieties of information with rich and varied structures and with compositional semantics, and non-communicative abilities to check whether some state of affairs matches a description, to notice a gap in information, expressed in a question) are also requirements for forms of information that are involved in perception, planning, expressing questions, formulating goals, predicting, explaining and reasoning.

From that viewpoint, communication between individuals in a public medium was a *secondary* function of language which evolved only after a more basic kind of competence evolved – the ability to use an internal, non-communicative, language in mental states and processes such as perceiving, thinking, supposing, intending, desiring, planning, predicting, remembering, generalising, wanting information, etc. and in executing intentions or plans.

It is clear that human children who cannot yet talk, and many animals that do not use an external language can perceive, learn, think, anticipate, have goals, threaten, be puzzled, play games, carry out intentions and learn and use facts about causation. I know of no model of how any of that can be done without rich information processing capabilities of kinds that require the use of internal languages with compositional semantics. But I know of no detailed model of what those 'prelinguistic' languages are, how they evolved, how they develop, how they work, etc.

Fodor, in [31], postulated a 'language of thought' (LOT) which was supposed to be innate, available from birth and capable of expressing everything that ever needs to be said by any human being, but he left most questions about how it worked unanswered and was not concerned with non-human animals. Moreover he supposed that external languages are translated into the LOT, whereas there is no reason to believe such translation is necessary, just as compiling to machine code is not necessary for computer programs to run, if an interpreter is available. Moreover if it were necessary, then substantive ontology extension during learning and development would be impossible.

Nobody is yet in a position to say what the prelinguistic languages do and do not share with human communicative languages. In particular, it may be the case that the main qualitative competences required for human language use already exist in these pre-linguistic competences in young children and other animals, and that the subsequent evolutionary developments related to human language were mainly concerned (a) with developing means of generating adequately articulated external behaviours to communicate their structures, (b) improving perception of such behaviours, and (c) extending internal mechanisms (e.g. short term memories) that are equally useful for sophisticated internal information processing (such as planning) and for communication with others.

If so, information (unconsciously) acquired, used and stored for later use in early childhood may have much richer structures and deeper semantic content than has previously been thought possible. Whether it includes the kind of content required to support psychoanalytic theories remains open.

Non-auditory communication

It is easier to achieve a large collection of perceptually discriminable signals by using independently movable fingers, hands, mouth, eyes, head than to do it all by modulating an acoustic signal – especially as that would interfere with breathing. So, if some animals already had hands for which they were learning to produce (and perceive, during controlled execution) large numbers of distinct manipulative competences concerned with obtaining and eating food, grooming, climbing, making nests, fighting, threatening, etc., then perhaps the first steps to communicative competence used structured movements, as in that involved producing and perceiving structured sign language, rather than vocal language.

There is much suggestive evidence, including the fact that humans find it almost impossible not to move hands, eyes, facial muscles, etc. when talking, even when talking on the phone to someone out of sight. More compelling is the ease with which deaf babies are able to learn a sign language, and the reported fact that some children with Down syndrome seem to learn to communicate more easily if sign language is used. Most compelling is the case of the Nicaraguan deaf children who invented their own highly sophisticated sign language, leaving their teacher far behind. See [32].[10]

An implication of this is that the human ability to develop linguistic competence does not depend on learning a language that is already in use by others. Language learning then appears to be a process of collaborative, creative problem-solving, which may be constrained by the prior existence of a social language, but need not be.

In [33] Arbib proposed that action recognition and imitation was a precursor to the evolution of language, but the arguments

[10] `http://www.indiana.edu/~langacq/E105/Nicaragua.html` 'A Linguistic Big Bang', by Lawrence Osborne *The New York Times* October 24, 1999. A five minute video including examples of the invented language is available here `http://www.pbs.org/wgbh/evolution/library/07/2/l_072_04.html`

are somewhat different, and do not share our hypothesis in [29] that the existence of a rich non-communicative language preceded the evolution of communicative language, though the commentary by Bridgeman [34] makes a similar point, emphasising pre-linguistic planning capabilities.

Once something like this form of structured communication had developed, enabling information that was previously only represented in internal languages to be shared between individuals via a public language, the enormous advantages of having such a communicative competence, both in solving problems that required collaboration and in accelerating cultural transmission, might have led to an unusually rapid process of selection for changes that enhanced the competence in various ways, e.g. developing brain regions specialised at storage of ever larger vocabularies and more complex rules for construction and interpretation of action sequences.

Another development might have been evolution of brain mechanisms to support construction and comparison of more complex symbolic structures, e.g. more deeply nested structures, and structures with more sub-structures.

The nature of linguistic communication

If linguistic communication evolved from collaborative non-linguistic activities, controlled by sophisticated internal languages, this must change our view of human language. It is often thought that linguistic communication involves a process whereby an information structure in one individual gets encoded in some behaviour which is perceived and decoded by another individual who then constructs the same information internally. However, in collaborative, creative, problem solving the shared physical and task contexts provide much of the information, and need not be communicated. All that is needed is whatever suffices to produce desired results, without copying an information structure from one brain to another.

If a mother hands a child a piece of fruit, the child may thereby be triggered to form the goal of eating it, without the mother having had the goal of eating it. Moreover, the mother does not need to anticipate the precise movements of bringing the food to the mouth nor the precise chewing and swallowing movements produced by the child. Those details can be left to the child. More generally, context allows communication to be schematic rather than concrete and detailed.

This is obvious when A asks 'Where are my keys?' and B replies 'Look behind you?'. B may not know exactly where the keys are except that they are on the table behind A. A turns round and sees the keys, getting the precise information he wanted, as a result of B's communication, even though B did not have that information. Both have achieved their communicative goals, but not by transmitting a specific information structure from B to A.

So, many kinds of linguistic and non-linguistic communication involve communication of a *partial* or *schematic* structure, leaving gaps to be filled by the receiver using information that may or may not be available to the sender. In that case, many linguistic expressions may be best thought of as having a higher-order semantics, to be applied to more specific non-linguistic information as context demands, in collaborative problem solving rather than in a process of information transfer. (This is related to Grice's maxims of communication, and to current concerns with 'situatedness' in language understanding, perception and action. `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0605`)

If the contents of linguistic communication *between* individuals make such use of the context then it may be equally true for thought processes *within* an individual. In that case, much of the information content of your mind is outside you. This may be one of the ideas people who emphasise the importance of 'embodiment' are getting at. However, everything said here could apply to minds in individuals with simulated bodies located in simulated environments. The important thing is what sorts of structures, processes and causal interactions occur within the individual's information processing architecture and how they are related to its environment. Whether that environment is physical or another virtual machine environment makes little or no difference to the kind of mind discussed here.

From sign language to sound language

Use of sign language has many problems. As has often been noted, it can work only when the sender's hands are free, and when the receiver can see the sender, thereby ruling out communication in many situations where visual attention must be used for another task, or where there are obstructions to vision, or no light is available. So perhaps our ancestors started replacing hand signals with sounds where possible. This could then lead in the usual way to a succession of evolutionary changes in which the vocal mechanisms, auditory mechanisms, and the neural control systems evolved along with physiological changes to the vocal tract, etc. But the ability to learn sign languages remains intact, even if not used by most people.

If all this is correct then it may have deep implications for clinical developmental psychology as well as for understanding precisely what the effects of various kinds of brain damage are.

VIII. Architectures that grow themselves

Much research in AI and robotics is concerned with what sort of architecture an intelligent system needs, and various rival architectures are proposed. If the architecture needs to grow itself, some of this research effort may have been wasted. It was argued in [3] that the architecture should not be regarded as genetically determined, but as grown using both innate and acquired meta-competences influenced by the physical and cultural environment, as depicted in Figure 1. So there may be considerable differences in how adult minds work, as a result of a different developmental trajectories caused by different environments. This might affect kinds of self-monitoring, kinds of self control, kinds of learning ability, and so on. Another implication is that attempts to quantify the influence of genes and environment in terms of percentages is completely pointless.

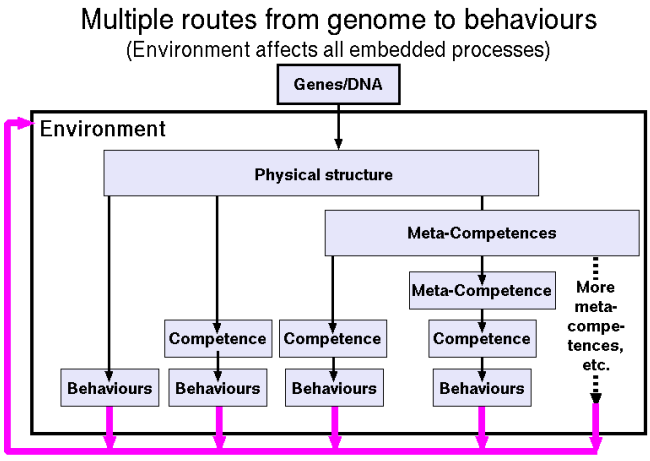

Fig. 1. This indicates some of the many and varied relationships between the genome and behaviours, produced at different stages of development after layers of competences and meta-competences have been built up. (From [3])

IX. Conclusion: machines in ghosts

Although I was asked to report on developments in AI that might be relevant to psycho-analysis I have instead focused on some very general features of requirements for AI models rather than specific AI models, because I think that despite all the advances achieved in AI, the models are still very primitive compared with human minds and have a very long way to go before they can be directly relevant to working therapists dealing with real humans.

Nevertheless, I hope it is now clear why a human-like machine, a human-like intelligent ghost, and indeed a human-like human must contain very specific kinds of information processing virtual machines. So our ability to understand human minds and ways they can go wrong, including acquiring false, beliefs, inappropriate motives, inadequate strategies, tendencies to be over-emotional, and worse, must be informed by deep knowledge about information-processing systems of the appropriate kinds.

The complexity of the machine, and the fact that self-monitoring mechanisms within it provide only a very limited subset of information about what is going on cause many individuals to start wondering what is really going on in them. Ignorance about information-processing mechanisms can make things seem so mysterious that some people invoke a special kind of stuff that is quite unlike physical matter. From there it is a short step to immortality, souls, etc., and the conceptual confusions discussed by Ryle in [35]. But if we don't go that far we may still come up with deep new ways of thinking about phenomena that were previously thought resistant to computer modelling. This could, among other things, lead to a revolution in psychoanalysis.

Acknowledgement

Thanks to Dean Petters, Jackie Chappell, Matthias Scheutz, Margaret Boden, members of the EU-Funded Cosy Project, and other collaborators and discussants over the years. Some of the ideas here were inspired by John McCarthy and Marvin Minsky, as well as Immanuel Kant, [6]. Related papers and presentations are in the Birmingham CogAff and CoSy project web sites. Jackie Chappell and Gerhard Pratl made useful comments on an early draft.

REFERENCES

[1] D. C. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press, 1978.

[2] A. Sloman, *The Computer Revolution in Philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press), 1978, http://www.cs.bham.ac.uk/research/cogaff/crp.

[3] J. Chappell and A. Sloman, "Natural and artificial meta-configured altricial information-processing systems," *International Journal of Unconventional Computing*, 2007, http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609.

[4] A. Sloman and J. Chappell, "The Altricial-Precocial Spectrum for Robots," in *Proceedings IJCAI'05*. Edinburgh: IJCAI, 2005, pp. 1187–1192, http://www.cs.bham.ac.uk/research/cogaff/05.html#200502.

[5] S. Harnad, "The Symbol Grounding Problem," *Physica D*, vol. 42, pp. 335–346, 1990.

[6] I. Kant, *Critique of Pure Reason*. London: Macmillan, 1781, translated (1929) by Norman Kemp Smith.

[7] A. Sloman, "The irrelevance of Turing machines to AI," in *Computationalism: New Directions*, M. Scheutz, Ed. Cambridge, MA: MIT Press, 2002, pp. 87–127, http://www.cs.bham.ac.uk/research/cogaff/00-02.html#77.

[8] A. Damasio, *Descartes' Error, Emotion Reason and the Human Brain*. New York: Grosset/Putnam Books, 1994.

[9] R. Picard, *Affective Computing*. Cambridge, Mass, London, England: MIT Press, 1997.

[10] A. Sloman and M. Croucher, "Why robots will have emotions," in *Proc 7th Int. Joint Conference on AI*. Vancouver: IJCAI, 1981, pp. 197–202.

[11] A. Sloman, "Towards a grammar of emotions," *New Universities Quarterly*, vol. 36, no. 3, pp. 230–238, 1982, http://www.cs.bham.ac.uk/research/cogaff/96-99.html#47.

[12] A. Sloman, "Real time multiple motive-expert systems," in *Proceedings Expert Systems 85*, M. Merry, Ed. Cambridge University Press, 1985, pp. 213–224.

[13] A. Sloman, "Damasio, Descartes, alarms and meta-management," in *Proceedings International Conference on Systems, Man, and Cybernetics (SMC98), San Diego*. IEEE, 1998, pp. 2652–7.

[14] A. Sloman, R. Chrisley, and M. Scheutz, "The architectural basis of affective states and processes," in *Who Needs Emotions?: The Brain Meets the Robot*, M. Arbib and J.-M. Fellous, Eds. Oxford, New York: Oxford University Press, 2005, pp. 203–244, http://www.cs.bham.ac.uk/research/cogaff/03.html#200305.

[15] A. Sloman, "Do machines, natural or artificial, really need emotions?" 2004, Invited talk: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cafe04.

[16] A. Sloman, "'Ontology extension' in evolution and in development, in animals and machines," University of Birmingham, UK, 2006, PDF presentation: http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0604.

[17] A. Sloman and R. Chrisley, "Virtual machines and consciousness," *Journal of Consciousness Studies*, vol. 10, no. 4-5, pp. 113–172, 2003.

[18] A. Berthoz, *The Brain's sense of movement*, ser. Perspectives in Cognitive Science. London, UK: Harvard University Press, 2000.

[19] A. Sloman, "How to derive "better" from "is"," *American Phil. Quarterly*, vol. 6, pp. 43–52, Jan 1969, http://www.cs.bham.ac.uk/research/cogaff/sloman.better.html.

[20] A. Sloman, "The mind as a control system," in *Philosophy and the Cognitive Sciences*, C. Hookway and D. Peterson, Eds. Cambridge, UK: Cambridge University Press, 1993, pp. 69–110.

[21] M. Lungarella and O. Sporns, "Mapping information flow in sensorimotor networks," *PLoS Computational Biolology*, vol. 2, no. 10:e144, 2006, dOI: 10.1371/journal.pcbi.0020144.

[22] A. Sloman, "Sensorimotor vs objective contingencies," University of Bimringham, School of Computer Science, Tech. Rep. COSY-DP-0603, 2006, http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0603.

[23] A. Sloman, "Architecture-based conceptions of mind," in *In the Scope of Logic, Methodology, and Philosophy of Science (Vol II)*. Dordrecht: Kluwer, 2002, pp. 403–427, (Synthese Library Vol. 316).

[24] A. Sloman, "Requirements for a Fully Deliberative Architecture," University of Bimringham, School of Computer Science, Tech. Rep. COSY-DP-0604, 2006, http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604.

[25] M. Minsky, "Matter Mind and Models," in *Semantic Information Processing*, M. Minsky, Ed. Cambridge, Mass,: MIT Press,, 1968.

[26] A. Trehub, *The Cognitive Brain*. Cambridge, MA: MIT Press, 1991, Online http://www.people.umass.edu/trehub/.

[27] N. Chomsky, *Aspects of the theory of syntax*. Cambridge, Mass: MIT Press,, 1965.

[28] E. Jablonka and M. J. Lamb, *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge MA: MIT Press, 2005.

[29] A. Sloman, "The primacy of non-communicative language," in *The analysis of Meaning: Informatics 5 Proceedings ASLIB/BCS Conference, Oxford, March 1979*, M. MacCafferty and K. Gray, Eds. London: Aslib, 1979, pp. 1–15, http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html.

[30] A. Sloman and J. Chappell, "Computational Cognitive Epigenetics (Commentary on [28])," *Behavioral and Brain Sciences*, 2007.

[31] J. Fodor, *The Language of Thought*. Cambridge: Harvard University Press, 1975.

[32] L. Osborne, "'A Linguistic Big Bang'," *The New York Times*, 24 October 1999, http://www.indiana.edu/ langacq/E105/Nicaragua.html. A five minute video including examples of the invented language is available at http://www.pbs.org/wgbh/evolution/library/07/2/l_072_04.html.

[33] M. A. Arbib, "From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics," *Behavioral and Brain Sciences*, vol. 28, no. 2, pp. 105–124, 2005.

[34] B. Bridgeman, "Action planning supplements mirror systems in language evolution," *Behavioral and Brain Sciences*, vol. 28, no. 2, pp. 129–130., 2005.

[35] G. Ryle, *The Concept of Mind*. London: Hutchinson, 1949.

Aaron Sloman Aaron Sloman studied for a BSc in mathematics and physics at Cape Town in 1956, then, after flirting for a while with mathematical logic was seduced by philosophy, and completed a D.Phil at Oxford in 1962, defending Kant's philosophy of mathematics. He later learnt that AI provided the best tools for doing philosophy and since about 1970 has been working on interdisciplinary problems concerned with understanding the space of possible designs for behaving systems including humans, other animals and possible future robots. In recent years this has particularly involved collaborating with a biologist studying cognition in birds and roboticists designing robots able to perceive and manipulate 3-D objects in the environment. The various strands of his work are summarised in `http://www.cs.bham.ac.uk/~axs/my-doings.html` He is a Fellow of AAAI, ECCAI and SSAISB, and was given an honorary DSc by Sussex University in 2006.