|

WHY PHILOSOPHERS SHOULD BE DESIGNERS

(Commentary on Dennett for BBS)

Aaron Sloman
1988

School of Cognitive and Computing Sciences
The University of Sussex

ABSTRACT

This is a short commentary on some aspects of D.C.Dennett's book *The Intentional Stance.* The paper criticises the "intentional stance" as not providing real insight into the nature of intelligence because it ignores the question HOW behaviour is produced. The paper argues that only by taking the "design stance" can we understand the difference between intelligent and unintelligent ways of doing the same thing.

Commentary submitted to
The Behavioral and Brain Sciences Journal
on:
D. C. Dennett
The Intentional Stance

Cognitive Science Research Papers
CSRP 97
School of Cognitive Sciences
University of Sussex
Brighton, BN1 9QN, England

Since 1991 Aaron Sloman has been at the School of Computer Science,
The University of Birmingham
http://www.cs.bham.ac.uk/~axs/

Why philosophers should be designers

Aaron Sloman

I agree with most of what Dennett says: most of the views he attacks are seriously flawed, often being either mere armchair pontifications about empirical and design issues or else ethical injunctions disguised as factual assertions.

Alas, there is also a subtle flaw in Dennett's own position which, if remedied, would enable his work to be more fruitful for the task of finding principles relevant to both the scientific study of existing intelligent systems (e.g. mice and men) and the design of new ones. This search for general principles is the central goal of AI: those who construe it as merely a branch of engineering ignore the content of much of the work in AI departments, journals and conferences.

Dennett's mistake is to separate the intentional from the design stance. He wants the intentional stance to focus entirely on rational behaviour and how to predict it, without regard to how the agent is designed, whether by evolution or engineers.

This resembles the attempts of 'phenomenalist' philosophers to construe physical objects solely in terms of their actual and possible perceptual manifestations. That failed because it is impossible to produce an adequate finite account of physical objects without using theory-laden concepts going beyond percepts. E.g. if only observables are allowed, but no reference to internal explanatory structures, then being made of iron has to be analysed in terms of *infinitely* many

'if <condition> then <consequence>'

statements. A similar pitfall awaits design-free intentionalism.

If concepts of mental states like 'desire' and 'belief' are to have the explanatory and predictive power required to cope with ordinary human creativity (and cussedness) then they must refer to states with generative power. In order to predict, or retrospectively explain, novel occurrences such as Fred's clever new tactics in a game of chess, Freda's diagnosis of the car engine failure, or Fido's ingenious method of opening the larder door, we must assume a collection of interacting internal states with recombinant causal powers. E.g. Freda noticed the similarity between the way this engine failed and an older one that had a fault in its distributor, and used her grasp of the similarities and differences to focus her search.

So new mental states are produced by the interaction of old ones, like understanding or producing a new sentence by re-combining old abilities to handle noun phrases, verb phrases, etc. We don't yet know much about how people do these things, apart from such general requirements as generativeness, and specific requirements as knowing the rules of chess or grammar, facts about arithmetic, and so on. (Whether such knowledge is conscious and accessible or not is irrelevant.)

We have only a few sketchy ideas about possible mechanisms. E.g. AI textbooks describe programs that parse sentences, analyse images and make plans, demonstrating the possibility of computational mechanisms with seminal forms of the required recombinant powers. These mechanisms possess, albeit in limited ways, indefinitely rich generative powers, enabling ever new states and interactions based on old ones. The set of sentences an AI program can in principle parse, the set of images a (good) AI program can in principle interpret, is infinite. Machine memory sizes and quantisation of measurements limit the sets, but the range of competence remains infinite in principle (though the *kind* of variation is limited, e.g. by the grammar rules).

Ordinary predictions and explanations of intelligent behaviour, including the attribution of rationality, assume a kind of design capable of supporting this infinite (though not unbounded) generative capacity. The precise nature of human mental states and their causal interactions remain unknown, but the requirement for generative mechanisms is clear. There are limits due to the finite size or speed of the brain, but those are different from limits based on non-generative designs. Moreover, the existence of a *culture*, with memory extensions such as books and computer files, extends the limits inherent in individual human brains.

Dennett suggests that this sort of thing is ‘a relatively noncommittal set of specs’ (Precis p. 7) However, it is crucially committal. Not all computational mechanisms can meet the design requirement. So it is matter of *fact* whether people do or not, not just a matter of taking up a stance that happens to be useful. For instance, a finite state machine with N states, K acceptable patterns of sensory input, and a decision table mapping current state (determined by previous history) and current input, into next state and next output, would lack the required generative power. Unlike a memory-limited recursive parser, this machine would not itself be able to use additional memory were it provided (though its designer could then extend the decision table). The limitations of the finite state machine are connected with the fact that at any one time it is in an indivisible state (state number 5996 for instance). An explicit decision table doesn’t have causally interacting sub-states, as we do and AI programs do. It doesn’t have the ability to create a novel state by building a new data-structure, as AI programs do.

The N-state K-input machine could function as a pre-compiled emulation of a truly intelligent system constructed to function efficiently in a particular carefully limited environment, giving all the appearance of a machine with beliefs, desires, planning abilities, etc. but totally lacking the ability to cope with any input not explicitly anticipated in the compilation process. Because of its design origins, though not its actual design, adopting the intentional stance will be a fruitful way to predict its behaviour in situations compatible with its decision table. If input pattern P turns up while it is in a state S and the table contains no entry for the pair (P,S) then its behaviour is undefined: it might just go mute. If by luck no such situation turns up the intentional stance will work. But that doesn’t make it *correct*, only *useful*, up to a point.

Apart from the requirement of genetic foresight, I suspect it would be physically impossible to provide the storage capacity required for a finite state table-driven human infant to cope as humans do for seventy or more years in any culture and almost any physical environment over a whole lifetime. The pre-compiled table has to support all the counterfactual conditional statements about what the person would have done had she grown up in a Tibetan temple, a soviet spaceship, a kibbutz on Mars, etc. I suspect many other animals (birds, mice, dogs, etc)

also have too much flexibility to be based on such finite state machines.

Some of the simpler organisms may be so lacking in generative capacity. Certainly not humans and chimps: evolution does not have sufficient foresight for such pre-compilation except for special reflexes. Most situations are dealt with by an economical and powerful generative mechanism based on separate sub-states involving motives, beliefs, preferences, stored skills, systems of concepts for describing new situations, a host of learning abilities, and so on. The generative mechanisms include the ability to create task-specific decision tables (e.g. trained reflexes, memorised dances and poems, etc.)

Exactly what internal states and causal powers are assumed when we use mentalistic language depends on the sophistication attributed to the agent. It's not always the same. Compare (a) an animal (or infant) whose every desire always tends to produce actions with (b) an agent who is able to notice conflicts of desires, or conflicts between desires and ethical principles, and choose between them. Case (b) requires a richer computational architecture supporting a wider variety of internal processes, with more kinds of temporary information stores, priority measures, etc.

Yet more architectural complexity is required if, instead of being fixed, the desires, 'desire generators' and 'desire comparators' are themselves all modifiable and extendable by higher level generators and comparators, as happens during moral and aesthetic education for example. (Sloman 1987).

Desire-like states as such therefore do not have a fixed set of causal powers: it all depends on the total architecture of the system in question. The same goes for every other type of mental state describable in ordinary language. Each different computational architecture is capable of supporting a different array of mental states. However they can ALL be compiled into one architecture: a finite state machine, if the machine is big enough and the set of inputs arbitrarily limited. But that doesn't mean the latter machine has the same set of states, despite behavioural indistinguishability over many lifetimes.

Because computational designs do not vary smoothly (you can't *smoothly* introduce a new sub-routine, or a meta-level desire comparator), there are sharp discontinuities in the space of possible designs (see Sloman 1985), and some of those discontinuities are to be found in evolution, contrary to Dennett's claim that 'there are no sharp discontinuities' (Precis p. 5). (Actually Darwinian evolution, unlike Lamarckian, *requires* discontinuities: only a finite number of discrete generations separates stages in evolution.) Probably many discontinuities occur in individual development too. A study of all these design discontinuities will enable us to make new distinctions between different kinds of mental capabilities.

Ordinary language will probably then evolve to track developments in scientific theory, as so often happens. For instance, before the advance of mathematics and physics, people could conceive of something speeding up or slowing down, but not the same thing increasing its velocity and decreasing its acceleration at the same time. Even now some people find that hard to grasp. Similarly, different types of mental states not currently distinguished become describable as a result of new understanding of design possibilities. For instance philosophers (including Dennett - Precis page 13) easily imagine a distinction between explicit and implicit stored

information, but not nearly as many distinctions between types of implicit store as computer scientists have been forced to invent in addressing design trade-offs between space, time, flexibility, generality, modularity, learnability, modifiability, debugability, etc.

Some who attack, and some who defend, computational designs postulate a fixed innate "language of thought" into which everything is to be translated (e.g. Fodor 1976). Dennett's critique of this view (Precis page 14) needs supplementing. If any kind of internal symbolism capable of being reasoned with is available, then translation into a language of thought is simply not needed because Carnapian meaning postulates (or a generalisation thereof) can extend a language with new untranslatable yet meaningful primitives. They have meaning insofar as the class of admissible models for the total set of 'sentences' is limited. But the meaning is never totally determinate. (For more on this see Sloman 1985 and 1986.) A related point is that programs in high level programming languages need not be compiled into machine code in advance of being run: e.g. LISP and BASIC programs are often stored in their original form and interpreted by other programs. Interpreted programs offer greater semantic flexibility since they don't require everything to be well defined in advance. Indeed, if the interpreter is itself interpreted it can be modified during execution according to need.

The semantic properties of an interpreter-driven machine can therefore be indeterminate in important ways. Likewise the mental states of human beings, if desires, hopes, beliefs, abilities, etc. are stored in forms that are interpreted as required in context rather than compiled in advance to some brain-language with a fixed semantics. This indeterminacy of intentional states is part of the basis of human creativity as shown in both science and art. For instance, it allows new concepts to crystallise: the core of all major scientific advance and human learning.

By skilful use of philosophical techniques of analysis to further the more detailed and systematic study of *design* requirements for various kinds of intentional abilities, Dennett could make a major contribution to our understanding of the mechanisms underlying intelligence, leaving less gifted philosophers to play with all those amusing puzzles about twin earthers, emigrating slot machines, and the like. The design stance makes some of the philosophical puzzles evaporate, because they are based on an over-simple view of the space of possible designs.

REFERENCES

Fodor, J.A., (1976) *The language of thought* Harvester press.

Sloman A, (1985) 'What enables a machine to understand?' in *Proceedings 9th International Joint Conference on AI*, Los Angeles.

Sloman A, (1986) 'Reference without causal links' in L. Steels, B. du Boulay, D. Hogg, editors, *Proc 7th European Conference on AI*, Brighton, North-Holland.

Sloman, A, (1987) 'Motives, Mechanisms and Emotions' in *Cognition and Emotion* Vol 1 no 3