NOTE:
A slightly shorter version of the following text was broadcast via net news
to the comp.ai and sci.philosophy.tech news group some time in 1988.
It began with this remark, which explains the closing question:

> I wasn't going to contribute to this discussion,
> but a colleague encouraged me.

A similar version appeared in *AISB Quarterly*, Winter 1992/3, Issue 82, pp.31-2

HOW TO DISPOSE OF THE FREE WILL ISSUE

Aaron Sloman

School of Computer Science

The University of Birmingham

Abstract:
Much philosophical discussion concerning freedom of the will is based on an assumption that there is a well-defined distinction between systems whose choices are free and those whose choices are not. This assumption is refuted by showing that when requirements for behaving systems are considered there are very many design options which correspond to a wide variety of distinctions more or less closely associated with our naive ideas of individual freedom. Thus, instead of one major distinction there are many different distinctions; different combinations of design choices will produce different sorts of agents, and the naive distinction is not capable of classifying them. In this framework, the pre-theoretical concept of freedom of the will needs to be abandoned and replaced with a host of different technical concepts corresponding to the capabilities enabled by different designs.

Philosophy done well can contribute to technical problems (as shown by the influence of philosophy on logic, mathematics, and computing, e.g. via Aristotle, Leibniz, Frege and Russell).

Conversely, technical developments can also help to solve or dissolve old philosophical problems. I think we are now in a position to dissolve the problems of free will as normally conceived, and in doing so we can make a contribution to AI as well as philosophy.

The basic assumption behind much discussion of freedom of the will is:

> (A) there is a well-defined distinction between systems whose choices are free and those whose choices are not free.

However, if you start examining possible designs for intelligent systems *in great detail* you find that there is no one such distinction. Instead there are many 'lesser' distinctions corresponding to design decisions that a robot engineer might or might not take – and in many cases it is likely that biological evolution tried both (or several) alternatives.

There are interesting, indeed fascinating, technical problems about the implications of these design distinctions. For example, we can ask how individuals with the different designs would fare in a variety of social settings, what they would be like to interact with, which sorts of tasks they would be able to achieve and which not. Exploring design details shows, I believe, that there is no longer any interest in the question whether we have free will because among the *real* distinctions between possible designs there is no one distinction that fits the presuppositions of the philosophical uses of the term "free will". It does not map directly onto any one of the many different interesting design distinctions. So (A) is false.

"Free will" has plenty of ordinary uses to which most of the philosophical discussion is irrelevant. E.g.

"Did you go of your own free will or did she make you go?"

That question presupposes a well-understood distinction between two possible explanations for someone's action. But the answer "I went of my own free will" does not express a belief in any metaphysical truth about human freedom. It is merely a denial that certain sorts of influences operated, such as threats or coercion by another person. There is no implication that *no* causes, or no mechanisms were involved. How could any lay person know that there are no causes, since we know very little about how our brains work?

The claim to have done something of your own free will simply illustrates a common-sense distinction between the existence or non-existence of particular sorts of 'external' influences on a particular individual's action. We could all list types of influences that might make us inclined to say that someone did not act of his own free will, some of which would, for example, lead to exoneration in the courts. But saying "I did not do it of my own free will because processes in my brain caused me to do it" would not be accepted as an excuse, or a basis for requesting forgivenness.

However there are other deeper distinctions that relate to different sorts of designs for behaving systems, but our ordinary language does not include terms for distinguish behaviour flowing from such different designs. Before we can introduce new theory-based distinctions, we need to answer the following technical question that lurks behind much of the discussion of free will.

> "What kinds of designs are possible for intelligent agents and what are the implications of different designs as regards the determinants of their actions?"

I'll use "agent" as short for "behaving system with something like motives". What that means is a topic for another day. Instead of *one* big division between things (agents) with and things (agents) without free will we'll then come up with a host of more or less significant divisions, expressing some aspect of the pre-theoretical free/unfree distinction. E.g. here are some examples of design distinctions (some of which would subdivide into smaller sub-distinctions on closer analysis):

- Compare (a) agents that are able simultaneously to store and compare different motives with (b) agents that have no mechanisms enabling this: i.e. they can have only one motive at a time.

- Compare (a) agents all of whose motives are generated by a single top level goal (e.g. "win this game") with (b) agents that have several independent sources of motivation (motive generators implemented in hardware or software), e.g. thirst, sex, curiosity, political ambition, aesthetic preferences, etc.

- Contrast (a) an agent whose development includes modification of its motive generators and motive comparators in the light of experience, with (b) an agent whose generators and comparators are fixed for life (presumably the case for many animals).

- Contrast (a) an agent whose motive generators and comparators change partly under the influence of genetically determined factors (e.g. puberty), with (b) an agent for whom they can change only in the light of interactions with the environment and inferences drawn therefrom.

- Contrast (a) an agent whose motive generators and comparators (and higher order motivators) are themselves accessible to explicit internal scrutiny, analysis and change, with (b) an agent for which all the changes in motive generators and comparators are merely uncontrolled side effects of other processes (as in addictions, habituation, etc.) A similar distinction can be made as to whether motives themselves are or are not accessible to explicit internal scrutiny, analysis and change.

• Contrast (a) an agent pre-programmed to have motive generators and comparators change under the influence of likes and dislikes, or approval and disapproval, of other agents, and (b) an agent that is only influenced by how things affect it. The former will be more likely than the latter to absorb the values of its culture.

• Compare (a) agents that are able to extend the formalisms they use for thinking about the environment and their methods of dealing with it (like human beings) and (b) agents that are not (most other animals?)

• Compare (a) agents whose motives are never inconsistent, e.g. because the latest motive always removes all others, and (b) agents that can simultaneously have incompatible motives (e.g. wanting to drink at the water hole and wanting not to go near that hungry looking lion crouching beside the water).

• Compare (a) agents that are able to assess the merits of different inconsistent motives (desires, wishes, ideals, etc.) and then decide which (if any) to act on with (b) agents for which motivator conflicts are always resolved using some automatic reaction, e.g. having a measure of 'strength' associated with each motive and always being driven by the strongest motive, or agents that are always controlled by the most recently generated motive (like very young children and perhaps some other animals?).

• Compare (a) agents with a monolithic hierarchical computational architecture where sub-processes cannot acquire any motives (goals) except via their 'superiors', with only one top level executive process generating all the goals driving lower level systems with (b) agents where individual sub-systems can generate independent goals. In case (b) we can distinguish many sub-cases, for instance:

(b1) the system is hierarchical and sub-systems can pursue their independent goals if they don't conflict with the goals of their superiors

(b2) there are procedures whereby sub-systems can (sometimes?) override their superiors (e.g. trained reflexes?)

• Compare (a) a system in which all the decisions among competing goals and sub-goals are taken on some kind of 'democratic' voting basis or a numerical summation or comparison of some kind (a kind of vector addition perhaps) with (b) a system in which conflicts are resolved on the basis of qualitative rules, some of which are determined genetically (from birth) and some of which are products of a complex high level learning system.

• Compare (a) a system designed entirely to take decisions that are optimal for its own well-being and long term survival with (b) a system that has built-in mechanisms to ensure that the well-being of others is also taken into account. (Human beings and many other animals seem to have some biologically determined mechanisms of the second sort - e.g. maternal/paternal reactions to offspring, sympathy, etc.).

• Compare (a) a system that includes some kind of random generator that determines some of its major decisions and (b) a system all of whose decisions are based on its motives, beliefs, preferences, etc. which in turn are produced by totally deterministic processes including long term learning.

• There are many distinctions that can be made between systems according to how much knowledge they have about their own states, and how much they can or cannot change because they do or do not have appropriate mechanisms. (As usual there are many different sub-cases. Having something in a write-protected area is different from not having any mechanism for changing stored information at all.)

There are some overlaps between these distinctions, and many of them are relatively imprecise, but all are capable of refinement and can be mapped onto real design decisions for a robot-designer (or evolution).

They are just some of the many interesting design distinctions whose implications can be explored both theoretically and experimentally, though building models illustrating most of the alternatives will require significant advances in AI e.g. in perception, memory, learning, reasoning, motor control, etc.

When we explore the fascinating space of possible designs for agents, the question which of the various systems has free will loses interest: the pre-theoretic free/unfree contrast totally fails to produce any one interesting demarcation among the many possible designs – though it can be loosely mapped on to several of them. However, different mappings will imply different implications for classifying an agent as free, or as unfree.

After detailed analysis of design options we may be able to define many different notions of freedom, with corresponding predicates:- free(1), free(2), free(3), ... However, if an object is free(i) but not free(j) (for i /= j) then the question "But is it really FREE?" has no answer.

It's like asking: What's the difference between things that have life and things that don't?

The question whether something is living or not is (perhaps) acceptable if you are contrasting trees, mice and people with stones, rivers and clouds. But when you start looking at a larger class of cases, including viruses, complex molecules of various kinds, and other theoretically possible cases, the question loses its point because it uses a pre-theoretic concept ("life") that doesn't have a sufficiently rich and precise meaning to distinguish all the cases that can occur. (This need not stop biologists introducing a new precise and technical concept and using the word "life" for it. But that doesn't answer the unanswerable pre-theoretical question about precisely where the boundary lies.)

Similarly "What's the difference between things with and things without free will?" may have an answer if you are contrasting on the one hand, thermostats, trees and the solar system with, on the other hand, people, chimpanzees and intelligent robots. But if the question is asked on the presumption that *all* behaving systems can be divided, then it makes the false assumption (A).

So, to ask whether we are free is to ask which side of a boundary we are on when there is no particular boundary in question, only an ill-defined collection of very different boundaries. This is one reason why it is that so many people are tempted to say "What I mean by 'free' is..." and they then produce different incompatible definitions.

In other words, the problem of free will is a non-issue. So let's examine the more interesting detailed technical questions in depth.

It is sometimes thought that the success of computational models of the human mind would carry the implication that we lack freedom because computers have no freedom. However, as I argued in section 10.13 of Sloman (1978), on the contrary such models may, at last enable us to see how it is possible for agents to have an architecture in which *their own* desires, beliefs, preferences, tastes and the like determine what they do rather than external forces or blind physical and chemical processes. This line of thinking is elaborated in the books and papers cited in the bibliography. Dennett (1984), in particular, analyses in considerable depth the confusions that lead people to worry about whether we are free or not.

Now, shall I or shan't I submit this.....????

Some possibly useful references

Beaudoin, Luc P. and Aaron Sloman, (1993) A Study of Motive Processing and Attention, in *Prospects for AI as the General Science of Intelligence, Proceedings AISB93*, IOS press, Amsterdam.

Dennett, D.C. (1984) *Elbow Room: the varieties of free will worth wanting* Oxford: The Clarendon Press.

Heckhausen, H. and J. Kuhl (1985), From wishes to action: The dead ends and short cuts on the long way to action, *Goal directed behavior: The concept of action in psychology,* M. Frese and J. Sabini, eds. Lawrence Earlbaum Associates: London.

Sloman, A. (1978) *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind* Harvester Press, and Humanities Press

Sloman A., and Monica Croucher, (1981) Why robots will have emotions, in *Proceedings 7th International Joint Conference on Artificial Intelligence* Vancouver, 1981, also available as Cognitive Science Research Paper 176, Sussex University.

Sloman, A. (1987), Motives, mechanisms and emotions, *Cognition and Emotion,* 1: 217-234.

Sloman, A. (in press), Prolegomena to a theory of communication and affect, *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues,* A. Ortony, J. Slack, and O. Stock, eds. Springer: Heidelberg, 229-260.