RESPONSIBLE COMPUTERS?

Yorick Wilks

Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003

ABSTRACT

The position paper argues that, on one coherent philosophical position, we can now say that computers have human attributes, and then go on to dicuss the route by which blame and punishment might be applied to them, and how they might be said to take on social obligations.

A. Human attributes and machines

It is a fact of common observation that people now anthropomorphise computers in their speaking and writing, and not only computers as such, but even their parts: "What the color chip is telling you is that it's in the background mode" a vision hacker said to me last week. That is no different from what we say of human wholes and parts, as in "my stomach is telling me it's lunchtime", and so such attributions do not, of themselves, have any consequences or relevance, legal or general.

But they are, nonetheless, a necessary **precondition** of any attribution of legal or other responsibility beyond the human pale. Sherry Turkle's recent book (1984) has given the sociological imprimatur, if it were needed, to the claim that usage is now like that, especially among small children.

More importantly, the fact that adults now talk and write that way has nothing to do with Turing tests and "being fooled by simulations", as some people acquainted with the historical AI literature might think: for the forms of words in questions are used by people who have never seen any plausible language or reasoning task performed by a machine, or rather have seen no such performances other than simulated fictions on TV and films. Given that TV viewers vastly outnumber computer scientists, it is those "performances" that are, I suspect, the driving forces behind the language changes under discussion.

But those changes themselves are perfectly real and, for anyone of a Dennetist tendency in philosophy (if I may use that word to refer to one who gives theoretical priority to successful explanatory vocabulary rather than underlying or direct ontological evidence: (Dennett 1978) machines may therefore now have certain key human characteristics. If that is so, then it may be the peg on which to hang any possible legal responsibility of machines or programs. But before turning to that, let me take a different case for comparison.

B. Dogs

In English common law, at least, there is already a well established and still operative precedent for a category of entities which are neither human, nor totally without responsibility. They are animals like dogs, which certainly pass the test of having appropriate attirbutions made to them, at least by a large part of the population. They are quite distinct from **ferae naturae** like tigers: if you keep a tiger and it does any wrong, you are responsible, for they are taken to be simple machines in your keeping. With dogs the situation is more complex and normally, though inaccurately, summed up in the cliche "every dog is allowed one bite"; the point being that a dog is not deemed savage simply because it bites someone once. It may, like us, be acting out of character. Whereas to be a savage dog is to be a habitual biter and in particular to have a savage character known to its owner. Tigers are not to be thought of as having characters to act out of: they are just machines that bite. This notion of having a character one could act out of is tightly bound up with the notions of moral and legal responsibility and blame.

Dogs are blamed and punished in analogous ways to people--in some countries both can be executed–and that is only because they share very similar (though importantly different) physiological structures. The problem with machines and their programs, even if we were to squeeze them into the same category as dogs, would be how to blame and punish them.

C. Responsibility

The difficulty can be avoided by always identifying humans, standing behind the machines and programs as it were, to carry the blame, in the sense in which there are always real humans standing behind agents and behind companies, which also have the legal status of non-human responsible entities ("anonymous persons" in much European law). In the case of companies with errant machines, the companies themselves (i.e. not their individual directors or shareholders) are responsible for a broad class of failures of their products and non-criminal actions by their agents, acting within the general futherance of company policy (see Lehman-Wilzig 1981). In those cases the punishment/destruction of machines and software packages would be merely a matter of internal company discipline and of no outside interest.

In most situations now imaginable, it will not be too hard to identify indivudals, if there is a need to do so, behind programs and machines. However, things may become more tricky as time goes on, and the simple substitution of responsible people for errant machines harder to achieve. There are two obvious possibilities here: first, there are already in existence enormous bodies of software, such as major bank and airline programs, that are the work of large numbers of individuals, that have been constantly edited and updated over many years, and are probably now without any adequate documentation. Those who could have written the documentation may well be dead. Such gigantic kludges function up to a point and it would be difficult and expensive to replace them. However, those who work with them are often unsure why they do what they do, or what they might do in the future. Errors committed by such software will be very hard to attribute to particular responsible individuals.

Secondly, it is a small step from that present reality to a future situation where we accord the machine itself greater authority over the state it is currently in than we now do to information gained from diagnostics, traces or even looking in its cabinet (see Wilks 1976). The complete print-out of the program run by such a machine may be horrendously long, unannotated and effectively

structureless. This situation can approximate as closely as you like to that of the human brain, where print-out is pretty useless, as far as establishing what "state" a person is in, and we tend, therefore, to give great authority, in courts and elsewhere, to what people say about their own states of mind, particularly for the attribution of a "guilty" state of mind, the **mens re**. That movement, through impenetrable software to ultimately inadequate diagnostics, is, I think, the progression by which blame (for machines) might creep in, despite the attempts by advocates of more perspicuous programming styles to keep it out.

D. Punishment

But what can we say of "machine punishment"? A machine can be turned off and smashed and the software will either go with it, or can be burned separately, provided we know we have **all** the copies! Only if some notion of computer blame had already crept in by the route I mentioned earlier, could we consider any of this destruction (or, more moderately perhaps, compulsory court-ordered edits to a program) as punishment. And then the issue for a court might be to decide whether to punish the software or the hardware, which would be in keeping with the speculations of the many philosophers who have toyed with the analogy hardware:software::body:soul-or-mind. But the weaknesses of that approach are well known by now in an era of machines almost hardwired for special software languages.

Anyone who finds something lacking in Joan of Arc's cry at the stake, that they were punishing her body but not her soul, will tend towards a position that persons are **embodied minds-or-souls,** and that perhaps only those can be punished, even in principle. It would then be a short step to a position that, if we were ever to talk of punishing intelligent machines, given that they could be blamed, it would have to be as machine-embodied software. It is a long way from the Lisp and Prolog machines of today, together with a little specialised speech and vision hardware, to a notion of a fully (and ineluctably) machine-embodied program. But that is the technical road we are going down, and it may also be the only one down which machine crime and punishment can possibly lie.

E. Obligation

In conclusion, let me return to the issue of "obligation" seen, as it were, from the machine's point of view: not just as a matter of "under what circumstances do we attribute responsibility, and hence blame, to machines?", which is what I have called the Dennetist question, but also as a matter of how would we introduce into programs the notion of "obligatory" or responsible action. This matter is far less speculative than the last, and one might say that current work in AI gives a fairly clear view of the way forward.

The issue is not just one of representations, as many AI issues are, but of cerrain actions by the machine being the acceptance of obligations, and marked internally as such. Searle (1969) set out bodies of rules for such notions as "acts of promising": conditions that must obtain, in terms of beliefs and goals, for a promise to have been made by an utterance. Versions of these rules have been programmed within AI, and have in certain ways improved upon Searle's work, particularly in establishing a clear notion of a hearer's/machine's computation of its own point of view of things, whereas his original rules are a mixture of speaker's and hearer's points of view.

What is worth noting here is that such work has normally been treated in AI as analyses of, say, "promising": as a linguistic mapping task from utterances such as "I'll give you $5 next week" to inner entities such as **PROMISE**. But what is often ignored is that Searle intended his work not as a linguistic task only, or even principally, but as an exploration of the foundations of moral obligation, i.e. of promising not "promising". One of the successful adaptations that Speech Act work has undergone in AI, rather than in linguistics or philosophy, has been to show the intimate connection betwen such analyses and planning theory. Such work could now go one step further towards Searle's original goal within the theory of obligation (whether or not he would concede it) by incorporating, within the planning aspects of Speech Act representations, the notion of actions deemed obligatory by a system for itself, and the tight connexion between such deeming and the external "social acts" that express the taking on of obligation e.g. "I, robot, swear...".

REFERENCES

[1] Dennett, D. Brainstorms, Bradford Books, Mass., 1978.

[2] Lehman-Wilzig, S.N. "Frankenstein unbound: towards a legal definition of Artificial Intelligence", Futures, 1981, pp.107-119.

[3] Searle, J. Speech Acts, Cambridge University Press, Cambridge, 1969.

[4] Turkle, S. The Second Self, Granada, London, 1984.

[5] Wilks, Y. "Putnam and Clarke and Body and Mind", Brit. Jnl. Philos. of Sci., 26, 1976, pp.213-225.