EMPOWERING AUTOMATIC DECISION MAKING SYSTEMS: GENERAL INTELLIGENCE, RESPONSIBILITY AND MORAL SENSIBILITY

Henry Thompson
Department of Artificial Intelligence
and
Programme in Cognitive Science
School of Epistemics
University of Edinburgh
Edinburgh EH8 9NW
SCOTLAND

0. Introduction

Before a human being begins to make decisions and take unsupervised actions in a professional capacity which significantly affect others, s/he is usually explicitly empowered to do so, by means of some socially and/or legally sanctioned process of training and evaluation.

At a time when it is being suggested that computational artifacts may take up roles with significant human impact, ranging from medical diagnosis to automatic launch on warning of nuclear missiles, it becomes appropriate to ask whether sufficient thought has been given to the question of establishing empowerment processes for such systems if they are to act autonomously without human supervision.

I believe that a careful and responsible investigation of this question will lead to a paradox - that the sorts of special-purpose, focussed systems which we can imagine being within reach technically will be manifestly and necessarily incapable of satisfying certain necessary criteria for empowerment, despite our inability to objectively define such criteria or design explicit tests to implement them. And this inability will in turn frustrate us if in the unforeseeably distant future we are finally in a position to build general-purpose, broadly intelligent systems*.

In what follows I consider first the proximate form of the paradox, as it applies to special-purpose systems, and then the longer term, more general case. The treatment is, given the constraints of space, time and the author's expertise, necessarily incomplete and anecdotal, rather than exhaustive and authoritative, but may at least serve to provoke debate.

1. Empowering Special-purpose Automatic Decision-Making Systems

In this section I am concerned with the kind of systems some at least among us appear to consider imminent - fully autonomous active decision-making systems designed for specific, fairly narrowly constrained tasks. In the near term we might imagine such systems arising by the closing of a sense-determine-act loop which to date still includes a human link, as in existing nuclear power reactor control systems, experimental disease diagnosis and treatment systems, and nuclear weapons command and control systems, the exact degree of automation of which we are not informed of. Or looking further ahead one might anticipate the automation of functions so far un-mechanised, ranging from bus driving to the administration of civil and criminal justice. I contend that no such systems should ever be empowered to act autonomously, because no test or procedure can ever be established which adequately establishes their competence.

1.1 The impossibility in some cases of realistic field testing

For an important subset of potentially empowered special-purpose automatic decision-making systems, realistic field testing is impossible owing to the intolerable cost of failure and/or the impossibility of creating the necessary test situation. Launch-on-warning systems are the most obvious example here, but any system concerned with quick response (thus eliminating the possibility of last-minute human intervention) to low-probability and/or low-frequency events will suffer from the same problem, in proportion to the cost of a wrong decision.

1.2 The inadequacy of testing under simulation

This problem is pervasive, and indeed defines in the end the class of empowerable special-purpose systems, namely, those for which exhaustive testing under simulation is possible. The problem with testing under simulation is that the test necessarily recapitulates the categorisation which underlies the specification of the system to be tested. Thus it cannot validate that categorisation. Even if we suppose that some combination of formal means of system specification, proofs of system 'correctness' based thereon and testing under simulation can (or will some day be able to) establish beyond doubt that a system implements its specification faithfully, we have still to validate that specification. In particular we have to validate the choices made in such a specification as to the dimensions of description relevant to the characterisation of the situations within which the system must act.

*I start from the assumption that no aspect of human intelligence and behaviour is in principle unachievable by a humanly constructed artifact. How long we shall have to wait for such artifacts, and whether their construction will incorporate any interesting insight into the mind, as opposed to the brain, are questions beyond the scope of this paper.

Consider the thermostat - an automatic decision-making system long since empowered. We are content with that empowerment not only because the cost of failure is acceptably low, and because physical law and demonstration convince us that, as per specification, contact is made or broken as a function of ambient temperature, but also because it is patent that the dimension of temperature is (almost always) the only one relevant to characterising situations sufficiently to determine whether they are 'furnace should be on' or 'furnace should be off.

1.3 The necessity of general intelligence

Why is it that one does not have to go far up the scale of complexity from thermostats before reaching a point where human supervision is uncritically assumed to be necessary, whether in existing systems such as automatic zero-visibility instrument landing systems for aircraft, or experimental diagnosis aids such as MYCIN? Not only from fear of system failure, I would claim, but also from intuitive appreciation of the potential inadequacy of specification. There is always a class of doubts expressed as "But what if..." which point to a dimension of significance omitted from the specification.

The distinction between a special-purpose system and one with general intelligence (e.g. human beings) is the ability of the latter to introduce into the decision-making process a characterisation of the situation along a normally irrelevant dimension. Unless one can convincingly demonstrate, as with the thermostat, that the dimensions of characterisation included in the special-purpose system include all those of conceivable relevance, impowerment is clearly inappropriate, indeed foolhardy. But for applications of sufficient complexity such demonstrations are unlikely to be possible. It is worth noting in this connection two instances of systems performing to specification, but incorrectly: The East Coast power failure of 1965 and the BMEWS alert of 1960 caused by radar echoes off the rising moon. It would seem, then, that any task of sufficient complexity which has an appreciable impact on humanity requires at least* quasi-human general intelligence to automate it safely.

II. Empowering Systems with General Intelligence

This option is much harder to come to grips with, since the construction of artifacts expected to exhibit general intelligence seems so much beyond us today. None-the-less some useful observations may be made. First of all, the paradox alluded to above is now clear- we may recognise that general intelligence is required in a system before it can be empowered to make a wide range of decisions autonomously, but how can we reliably determine that a candidate for empowerment *has* it? The Turing test in its various forms may be adequate in the intellectual or academic spheres, but, not to put too fine a point on it, would you bet your life on it? The ability to recognise and accomodate to the unexpected is almost by definition not susceptible to reliable test. It is instructive to consider how this issue is dealt with in empowering human beings. Interestingly enough to a large extent it isn't. We appear to take it for granted, in the established processes leading to the empowerment of doctors, judges, pilots, nurses, teachers etc, that the candidates are possessed of the non-specialist human ability to be appropriately sensitive to any and all relevant aspects of the context of the decision-making situation. To the extent that the question arises, it appears to be confronted obliquely and informally, rather that as an explicit part of the empowerment process.

Before confronting the answer to our problem which this observation points to, a partial diversion is in order, to consider the further criteria for empowerment which emerge when we imagine perhaps the most extreme possible case, that of a fully autonomous empowered decision-making system dispensing criminal justice.

II.1 Responsibility and moral sensibility

It seems to me that before we would consider empowering anything to sit in judgement over ourselves and our fellows, we would demand above and beyond the above-mentioned general intelligence, to say nothing of demonstrated legal competence, a recognition of the responsibility entailed by the role of judge. I am no theorist or philosopher of law, but it seems clear to me that despite what we hear about the justice system being the rule of law, not of men, we none-the-less count on a good judge's humanity to temper justice with mercy, to be unavoidably influenced by that which s/he shares with those brought before him/her. The responsibility which a judge bears for his/her decisions influences those decisions in a crucial, albeit ineffable, manner. But to admit this is to admit as relevant to our concerns the question of the nature of 'humanity', considered as a quality rather than a tautological property of *homo sapiens.* Now the reason for this diversion into matters judicial is I hope clear - determining general intelligence is only a sub-part, a rather small part, of determining humanity. If we assume rather uncritically, on the basis of indirect, subjective evidence, the generalised plasticity of intelligence of human candidates for empowerment, how much more uncritically and implicitly we assume their humanity!

*This is not to rule out the possibility that there are tasks *no* system can perform. Whether *any* system, human, artifact or hybrid, can rationally be required to decide whether or not an enemy missile attack is underway and to launch missiles in reply, all within eight minutes, seems unlikely at best. On the other hand it is clearly a moral and political decision what level of risk is tolerable in return for the benefits of automation. In the case of the power grid, with probability of successful operation reasonably high, based on past performance, and cost of failure, although high in inconvenience, likely to be low in terms of human lives, the risk (the integral of probability times cost) is probably worth the benefits. In the case of launch-on-warning, with probability of failure high, owing to the afore-mentioned impossibilities of effective testing, and cost of failure enormous, the risk is intolerable.

II.2 The only reliable test for humanity

In the end, then, I am led to suggest that the only test we could ever sensibly trust before empowering an automatic decision-making system is the one we subject human beings to: they will have to pass as human in the course of ordinary life. The test for humanity is being able to successfully participate in the human form of life, to convincingly *da-sein.*

HI.Conclusions: On Spirituality and Hybrid Systems

One thing that follows from the preceding line of argument is that the current disinclination, to put it mildly, of Artificial Intelligence and Cognitive Science to treat the spiritual side of human nature seriously is a grave mistake. For if recognition of responsibility arising from moral sensibility has a causal role to play in human decision making and human behaviour more generally, then the origin of moral sensibility in man's spirituality becomes a necessary subject of study. The fact that concern with the twin questions of Why is there something rather than nothing' and 'How ought I to live my life' is symptomatic of the human condition is of as much ultimate significance to theories of mind as are the nature of syllogistic reasoning or mental representations of grammar, and it may be that postponing an investigation of the essense of spirituality in favour of the current exclusive investigation of the essense of rationality may render the whole enterprise literally incoherent.

On a more practical note, if one concludes from the first section that for the foreseeable future in all systems of any consequence we must keep people in the loop, our problems don't disappear. Keeping people in the loop - building hybrid systems - is not as easy as it sounds, either to require or to do. If the human participation in a hybrid system is reduced to pushing a button in response to a light, no useful supervision has been accomplished. And how to design a genuinely hybrid system which *does* provide effective supervision is an open question. The experience of the Three Mile Island disaster suggests the we are a long way from being able to build systems which effectively integrate human beings' general intelligence with computers' special-purpose expertise to produce an ensemble capable of flexible and informed responses in high-pressure situations.

If the air of this talk has seemed overly dour and pessimistic, I think this is a necessary antidote to the facile optimism of too many of our more visible representatives in the media. It is our responsibility, as scientists and as human beings, to do our best to see that such optimism is balanced by an informed and skeptical realism before the inevitable social (and mortal) cost has to be paid.

ACKNOWLEDGEMENTS

Many of the ideas in this talk were developed in discussions at meetings of the Edinburgh Computing and Social Responsibility group, and with Anne Lee, Brian Smith, Susan Stucky and Catharine Thompson, to all of whom my thanks.