29

Ethical machines

I. J. Good
Virginia Polytechnic and State University
Blacksburg, USA

The notion of an ethical machine can be interpreted in more than one way. Perhaps the most important interpretation is a machine that can generalize from existing literature to infer one or more consistent ethical systems and can work out their consequences. An ultra-intelligent machine should be able to do this, and that is one reason for not fearing it.

INTRODUCTION

There is fear that 'the machine will become the master', especially compounded by the possibility that the machine will go wrong. There is, for example, a play by E. M. Foster based on this theme. Again, Lewis Thomas (1980) has asserted that the concept of artificial intelligence is depressing and maybe even evil. Yet we are already controlled by machines – party political machines.

The urgent drives out the important, so there is not very much written about ethical machines; Isaac Asimov wrote well about some aspects of them in his book *I Robot* (1950). Many are familiar with his 'Three Laws of Robotics' without having read his book. The three laws are:

"1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

Originally, I thought the three laws were mutually incompatible because they are not quantitative enough, but I found that Asimov had not by any means overlooked the quantitative aspect.

In one chapter of the book a robot on another planet refuses to believe that men, inferior as they are, can construct robots, and it also does not believe that Earth exists. Nevertheless the robot has religious reasons for keeping certain pointer readings within certain ranges, and it thus saves Earth from destruction. Thus the robot does not violate the first law after all. I was unconvinced by this idea, but it does suggest the possibility of a robot's being largely controlled by its 'unconscious mind', so to speak, in spite of misconceptions in its 'conscious mind', that is, by the operations handled by the highest control element in the robot.

Later in the book, so-called 'Machines', with a capital M, are introduced that are a cut above ordinary robots. They are ultra-intelligent and are more or less in charge of groups of countries. A subtle difference now occurs in the interpretation of the first law which becomes (p. 216) "No machine [with a capital M] may harm humanity; or, through inaction, allow humanity to come to harm". And again "...the Machine cannot harm a human being more than minimally, and that only to save a greater number".

Unfortunately it is easy to think of circumstances where it is necessary to harm a person very much: for example, in the allocation of too small a number of dialysis machines to people with kidney disease.

Asimov's book has the important message that intelligent machines, whether they have an ordinary status or are ultra-intelligent presidents, should be designed to behave as if they were ethical people. How this is to be done remains largely unsolved except that the flavour is utilitarian.

The problem splits into two parts. The first is to define what is meant by ethical principles, and the second is to construct machines that obey these principles.

ETHICS

The problem of defining universally acceptable ethical principles is a familiar unsolved and possibly unsolvable philosophical problem. If this problem could be solved in a fully satisfactory manner, then the problem of constructing a machine that would obey these principles would not be difficult. For, in a known parody of Wittgenstein (Good 1976), we may say that

> Was sich überhaupt sagen lasst
> lasst sich klar sagen
> und es lasst sich programmeirten sein.

[That is, "What can be said at all can be said clearly, and it can be programmed".]

The programming of ethics was initiated by early philosophers. According to Abelson (1967, p. 82), "Ethical philosophy began in the fifth century BC, with the appearance of Socrates, a secular prophet whose self-appointed mission

was to awaken his fellow men to the need for rational criticism of their beliefs and practices".

The article points out that Greek society at the time was changing rapidly from an agrarian monarchy to a commercial and industrial democracy. People were given power who, in Abelson's words, "needed a more explicit and general code of conduct than was embodied in the sense of honour and *esprit de corps* of the landed aristocracy". Similarly today's society is changing rapidly, and the machines that may gain power will also need a more explicit formulation of ethical principles than the people have who now wield power.

Unfortunately, after 2500 years, the philosophical problems are nowhere near solution. Do we need to solve these philosophical problems before we can design an adequate ethical machine, or is there another approach?

One approach that cannot be ruled out is first to produce an ultra-intelligent machine (a UIM), and then ask it to solve the philosophical problems.

Among the fundamental approaches to ethics are utilitarianism, contractualism (see, for example, Rawls 1971, who however, does not claim originality), and intuitionism, and various shades and mixtures of these approaches.

I tend to believe that the UIM would agree with the Bayesian form of utilitarianism. The Bayesian principle of rationality is the recommendation to "maximize expected utility', that is, to choose the act that maximizes $\Sigma p_i u_i$, where the u_i's are the utilities of various mutually exclusive outcomes of some potential action, and the p_i's are the corresponding probabilities. This principle is to some extent a definition of "utility", but it is not a tautology; it is more a principle of consistency. The development of the neo-Bayes-Laplace philosophy of rationality by F. P. Ramsey (1931), and L. J. Savage (1954) amounts to this: that a person or group that accepts certain compelling desiderata should act as if he, she, or it had a collection of subjective probabilities and utilities and wished to maximise the expected utility.

The social principle of rationality presents various difficulties:

(i) The estimation of interpersonal utilities if these are to be added together.

(ii) The question of whether the whole world (or galaxy etc.) should be taken into account with equal weights assigned to all people (or beings) or whether each society and individual should give much greater weight to itself, possibly in the hope that Adam Smith's "hidden hand" would lead to global optimization.

(iii) The assignment of weights to future people. Should the future be discounted at some specific rate such as 1% per year? The more difficult it is to predict the future, the higher the discounting rate should be.

(iv) The assignment of weights to animals. Should the weight given to any organism be some increasing function of degree of awareness? Should we assume that machines or even animals or slaves are zombies with no awareness and therefore have no rights?

One interpretation of ethical behaviour by a person is behaviour that tends to maximize the expected utility of a group to which he belongs, even if he suffers by so doing.

More generally an ethical problem arises when there is a conflict of interest between one group G and another, G', where a group might consist of only one person, and where the groups might intersect and one of the groups might even contain the other. It is possible too that one of the groups consists of people not yet born, or it might consist of animals. G might be one person, and G' the same person in the future. For example, we might criticize a machine for turning itself on if we believe that this would cause it damage.

I have been expressing in unemotional language the basis of many dramatic situations. For example, in *The Day of the Jackal*, de Gaulle's life is saved by the French Secret Service who obtained vital information by means of torture. Was this justified? Should we praise the brave German soldiers who laid down their lives for the sake of a criminal lunatic?

When a person acts rationally he uses his own utilities. If society is perfectly well organized the person will perform the same acts whether he uses his own utilities or those of the society. If a person seems to sacrifice his more obvious advantages for the sake of other people, then those other people would call him ethical. This would sometimes be because the interests of others are built into his personal utilities, and sometimes indirectly out of long-term self-interest.

Some people and some societies put more or less emphasis on different aspects of the Good, such as honesty, duty, love, loyalty, kindness, humility, religious feeling, bravery, and fairness or justice. The utilitarian regards all these aspects as derivative. For example, justice is regarded by the utilitarian as a useful concept because it makes a scheme of incentives more credible and so encourages legal, and perhaps ethical, behaviour. Similarly the justification of loyalty is that it encourages the leaders to be benign, and the main objection to terrorism is that it increases the probability of a ruthless Government. If a completely formalized mathematical theory of utility could be produced, then these derivative concepts would emerge in the form of theorems.

It might seem that a utilitarian must believe that the ends justify the means. Although he would certainly recognize the relevance of outcomes of acts, as would even most intuitionists, he might still agree, for example, with Aldous Huxley that the means are likely to *affect* the ends.

Possible Meanings for an Ethical Machine

In a sense, any machine, such as a pocket calculator, in good working order, is ethical if it is obedient. A slightly more interesting example is a homing missile because it has a little intelligence and is more like a kamikaze. Obedience by a person to the terms of a contract can certainly involve ethics, and obedience is also a quality that enables prisoners to earn remission of sentence, but it is not much of a criterion by itself. After all, most mobsters and Nazis are or were

obedient, so we need something more than obedience before we can feel happy about calling a machine ethical.

Another interpretation of an ethical machine is one that helps a person to be ethical by fairly straightforward information retrieval. Examples of such machines or programs, are:

(i) A machine that retrieves legal information. This enables an attorney to defend his client, or a judge to decide on sentences similar to those given in the past. Some judges have been guilty of exceedingly unethical behaviour, amounting almost to murder, through not having this kind of information or perhaps by pretending that they did not have it.

(ii) A machine that retrieves medical information.

Warren McCulloch (1956) defined an "ethical machine" as one that learns how to play a game by playing, but without being told the rules. He finishes his article by describing a man as "a Turing machine with only two feedbacks determined, a desire to play and a desire to win".

My concept of an ethical machine is somewhat different. I envisage a machine that would be given a large number of examples of human behaviour that other people called ethical, and examples of discussions of ethics, and from these examples and discussions the machine would formulate one or more consistent general theories of ethics, detailed enough so that it could deduce the probable consequences in most realistic situations.

As an example of this kind of machine or program let us consider the implicit utilities of medical consultants. This example was discussed by Card & Good (1970). The idea is that a team of medical consultants is to be asked what decisions they would make under various circumstances. These circumstances are defined by a set of indicants, and the probabilities of various outcomes are to be estimated independently of the decisions of the consultants. The probabilities and the decisions form the data for the calculations. A complication is that there might be inconsistencies in the decisions. It should be possible then, by an algorithm described in the article, to infer the implicit utilities that the consultants assign to the various outcomes. I don't know whether the algorithm has yet been applied in practice. It was not part of the investigation to assume different scales of fees to be paid to the consultants, nor to examine the effects of malpractice suits.

Concluding Remarks

A somewhat similar investigation has been carried out by Jones-Lee (1976) concerning the value of human life. (See also, for example, Mooney 1970.) The point of such investigations is to help decision-making in connection with, say, road safety. Some people object to such calculations on the grounds that life is priceless, overlooking that money saved on road safety can be spent, for example, on hospitals. It seems, however, inescapable that computer aids to administrative decision-taking will soon of necessity be performing calculations of this nature. So the problem of the ethical machine is already upon us.

REFERENCES

Abelson, R. & Nielsen, K. (1967). "Ethics, history of" in *The Encyclopedia of Philosophy*, vol. *3* (New York: Macmillan & The Free Press), 81–117. (Abelson wrote the part up through the 19th century.)

Asimov, Isaac (1950). *I Robot* (Garden City, New York: Doubleday, 1963 edn.).

Card, W. I. & Good, I. J. (1970). "The estimation of the implicit utilities of medical consultants", *Mathematical Biosciences* 6 (1970), 45–54.

Good, I. J. (1976). Pbi #322 in "Partly-baked ideas", *Mensa Journal International*, No. 193 (Jan. & Feb.), p. 1.

Jones-Lee, M. W. (1976). *The Value of Life: an Economic Analysis* (London: Martin Robertson).

Mooney, G. H. (1970). "The value of life and related problems", U.K. Ministry of Transport; mimeographed, 84 pp.

Ramsey, F. P. (1931). "Truth and probability" (1926), and "Further considerations" (1928) in *The Foundations of Mathematics and Other Logical Essays* (London: Kegan Paul).

Rawls, J. (1971). *A Theory of Justice* (Cambridge, Mass.: Harvard University Press).

Savage, L. J. (1954). *Foundations of Statistics*, New York: Wiley (2nd edn; Dover Publications, 1972).

Thomas, L. (1980). "Notes of a biological watcher. On aritificial intelligence", *New England J. Medicine* (Feb. 28), 506–507.