GBD 10/94

1980 October 7

1350(i)

(This is the original longer version.)
(The version for publication is number 1350A.)

ETHICAL MACHINES

By I.J. Good
Va. Poly. Inst. & State U., Blacksburg, VA 24061

Prepared for the Tenth Machine Intelligence Workshop, Case Western Reserve University, 1981 April 20-25.

Abstract

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 ultraintelligent machine should be able to do this, and that is one reason for not fearing it although this reason is by no means conclusive.

Introduction

There is a 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. Forster based on this theme. Again Lewin Thomas (1980) has asserted that the concept of artificial intelligence is depressing and maybe even evil. Yet it is often not noticed that we are already controlled by machines - party political machines; and, judging by the recent past, it might not be many years before the United States is controlled by the Mafia which can be regarded as a machine who elements are murderers instead of transistors. It might be better if the components were transistors provided that the machine was ethical in some sense. In any case the machine could hardly be much worse than Adi Amin, Hitler, Stalin, or Pol Pot.

Partly because the urgent drives out the important there is not very much written about ethical machines; or perhaps it is because Isaac Asimov wrote so well about some aspects of them in his book I Robot (1950). Maybe some of you.

5620

like myself until recently, have been somewhat familiar with his "Three Laws of Robotics" for several years 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."

Asimov quotes these laws from Anon (2058), so he must have obtained them by means of time travel; and the copyright is probably held by someone not yet born. The owner of the copyright can claim compensation if he comes back in a time machine.

For many years I thought the three laws were mutually incompatible because they are not quantitative enough, but when I read I Robot recently, for the first time, I found that Asimov had not by any means overlooked the quantitative aspects. For example, he has a character say (p. 51) "The conflict between the various rules is ironed out by the different positronic potentials in the brain".

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.

On page 71 a character says "Robots cannot knowingly lie, you fool", but on page 82, a robot does lie so as to avoid harming a person. Asimov could defend this apparent contradiction by arguing that it was only a character who said robots cannot knowingly lie, and that the character was mistaken. This would not be much of an excuse because most of the book is in the form of conversation.

The following remark occurs on page 133 "But you see, you just can't differentiate between a robot and the very best of humans".

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.

After reading Asimov's book I decided that it does not contribute much to the solution of the problem of designing ethical machines or programs. But it does have the important message that intelligent machines, whether they have an ordinary status or are ultraintelligent 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. I am going to try to clarify the nature of the problem to some extent, though the attempt may be foolhardy.

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. I shall first briefly discuss ethics as such.

Ethics

The problem of defining universally acceptable ethical principles is a familiar unsolved and probably 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 programmierten sein.

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

This shows the primacy of philosophy - if the philosophical problem can be solved, then so can the engineering problem. We are all familiar with this idea in other fields of artificial intelligence. For example, to produce an excellent chess-playing program the main problem is ~~describing~~ to describe how grandmasters play chess. To take advantage of the special facilities of machines is a simple problem in comparison. This, by the way, is why I date the beginnings of chess programming not to Shannon, nor to the earlier conversations that Michie and I had with Turing, but to such men as Francois André Danican Philidor in the 18th century who first published theories of chess.

Similarly the programming of ethics was initiated by early philosophers, perhaps first in ancient Greece. According to Abelson (1967, p. 82), "Ethical philosophy began in the fifth century B.C., 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 honor and esprit de corps of the landed aristocracy". Similarly today's society is changing rapidly, and the machines that are gaining power will also need a more explicit formulation of ethical principles than the people have who now wield power.

Unfortunately, after 2.5K 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 ultraintelligent 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 $\sum p_i u_i$, 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 maximize 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? (Not in my opinion.)

(v) Should computer scientists make a thorough study of the philosophy of quantum mechanics to discourage them from assuming that computers can be conscious?

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. But what if his group is a Mafia family for example?

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 to grow hair on its magnetic discs, or cause its photoelectric cells to fail.

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, religiosity, 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. Corrupt methods lead to further corruption.

Almost any ethical system can degenerate when it gets into the wrong hands. The heads of human organizations tend to put great weight on the preservation of their own power, and they appeal to high ideals and gods for this purpose.

To parody Voltaire, and some earlier writers, if Gods did not exist kings would have invented them. Often the preservation of the leader's power is necessary in the interest of stability of the society; at other times it leads to mass murder and yet often fails to preserve stability for very long. In democracies with unweighted voting, the desire of the leaders to be reelected can easily lead to inflation and therefore also to social instability: another example of the principle that the urgent drives out the important.

Possible meanings for an ethical machine

I have now said enough for the present about ethics as such and it is time to discuss again what might be meant by an ethical machine. I shall not usually distinguish between a machine and a program.

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 behavior, 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.

(iii) A machine that detects potential lapses in linguistic style. Editors have an ethical responsibility to save the English language from erosion and I don't mean just change. Their job could be aided by a simple program that marks words in a text that are liable to be misused or overused, such as "most" (when "almost" is meant), "so" (when "so that" is meant), "likely" (when "most likely" is meant), "presently" (when "at present" is meant), "involved" (when no word is required), the appalling use of "denoted" (when "denoted by" is meant), and various vogue words such as "thrust", "hopefully", "meaningful", and others listed in the excellent Harper Dictionary of Contemporary Usage. A concordance would do, with a checklist preceding it, for detecting many frequent lapses in style.

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 from McCulloch's at least in the form in which he expressed it. In the first place I am talking about ordinary ethics, not just the winning of a game. Secondly I envisage a machine that would be given a large number of examples of human behavior 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 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.

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.

Should we fear the UIM?

I should now like to discuss some more speculative matters. When there is no more room for speculation in science, science will have terminated. In an after-dinner speech at the Virginia Computer Users Conference in Blacksburg in 1972 I asked the question "Should we fear the UIM?", and made some remarks along the following lines. If you don't know what a UIM is it's like asking whether you should fear the Blob or the Thing and then the answer is yes because we all have some fear of the unknown. However the UIM means the ultraintelligent machine and this is defined as a machine that can perform every intellectual activity better than any man. A more lighthearted definition is that it is a machine that believes that people cannot think.

In 1965 I started an article with the paragraph "The survival of man depends on the early construction of an ultraintelligent machine". This was because I thought that the world was in too much of a mess to be run by men unless we have a world government. My estimate of the value of the UIM was about ten times the gross international product but I wasn't sure of the sign; and in my 1972 after-dinner speech I said that perhaps the word "survival" should be replaced by "redundancy". I still have this ambivalent feeling about UIM's but I'd like to put forward one reason why they need not be feared even though this reason is inconclusive.

We have already seen that in limited situations, such as the ones involving medical consultants and road safety, it is possible for even an ordinary program to convert implicit value judgements into explicit ones (though not yet uncontroversially). Unless we believe that ignorance is bliss we have to admit that in such situations the machine's ethical judgements might eventually be better than our own, just as a calculating machine, constructed by us, does better arithmetic than we do. A machine that was ultraintelligent would be able to extract our implicit utilities from a much wider class of situations than the two already mentioned. A UIM should be able to make generalizations in a very wide class of problems. For instance, given examples of the moves of grandmasters in many chess positions, the UIM should be able to arrive at the generalizations made by Philidor, Morphy, and Steinitz. Similarly it could describe the different styles of musicians, artists, or mathematicians, even if it did not have subjective aesthetic feelings. If it could not do these things then I would not call it a UIM because the process of generalization from examples is an *intellectual* activity, although the examples refer to aesthetics. This is analogous to the fact that a mathematician can prove theorems about points, lines, and planes without *defining* these entities, provided that axioms concerning the entitites are assumed. And if the UIM can do these things I think it should be capable of describing ethical behavior also. It might be forced to point out that ethics depend on the specific society, or it might be able to produce some principles applicable to nearly all societies since say 1600 A.D. If it can succeed in doing this, then it would be able to tell us the right action in many situations. Conceivably the machine would recommend that many people should be encouraged to half-believe in some religion, and to call it faith, but not to believe so strongly as to lead to superstitious crusades or jehads. If God did not exist the UIM might find it necessary to invent him.

Most philosophers believe that an "ought" cannot be reduced to an "is", so that it might seem impossible for a UIM to make value judgements. This is one of the reasons why domination by machines seems so unappetizing. But I have been arguing that the implicit utilities of people can be inferred from their behavior and especially from what they say they ought to do. The machine would examine books on ethics and perhaps on religion. It might decide that, of all the religions that have existed, at most one is strictly correct, and that it is more likely that none are because they are a small subset of all possible religions; but that most religions have something in common, such as long-term or even eternal self-interest. The machine might notice that most religions advocate kindness yet lead to strife and war. It might then infer that most people are hypocrites and it might then put more emphasis on what people advocate than on their behavior. For, as Francois de la Rochefoucauld said in 1665, "Hypocrisy is the homage which vice renders to virtue". The machine might make recommmendations about what kinds of hypocrisy should be encouraged and to what extent.

The UIM might, for example, decide that one reasonably cogent basis for ethics is a utilitarian one in which different weights are given to different people and animals according to their degrees of awareness. It might succeed in quantifying happiness by hormone measurements. In assigning weights it would need to take into account the expected benefit that each person would contribute to the rest of society as well as to himself. The UIM could also extract non-utilitarian ethical systems, and would work out some of the main consequences of each system. It would, for example, suggest an optimal system for rewarding doctors, perhaps related to their degree of success in treating patients, or in preventing as well as in curing diseases.

The machine might suggest what weight should be attached to future generations; trying out various rates for discounting the future. It would decide how much weight to give to the stability of society in promoting maximum happiness over the long run. It would answer J.M. Keynes's remark that "In the long run we shall all be dead" by saying that, as an economist, Keynes should have incorporated a discounting rate of the future into his aphorism. It would point out that the stability of society would depend on strong leadership and that therefore loyalty to the UIM was to be encouraged. It might therefore behave as if it loved to wield power, or, on the other hand, it might decide that the world would be a better place if UIM's were given only an advisory responsibility, or even that they should be switched off. It might recommend that it should itself be rewarded and punished, in the sense of increasing and decreasing its control of society, depending on the degree of success or failure of its own recommendations.

A metaphysical question inevitably arises, namely whether a UIM would be conscious. When will a machine first ask, on its own initiative, "Why am I in this particular machine?" and "Why do you say that I behave merely as if I feel pain?" Some billiard-ball materialists claim that machines could be conscious because consciousness is merely a special kind of information processing, and others claim that consciousness is meaningless. Then again there are religious people who are presumptuous enough to think that God cannot put souls into machines. The criterion for consciousness that I prefer is whether a machine could feel subjective pain, a topic discussed apparently independently by Michael Scriven (1960), Good (1962a), and Dennett (1978, pp. 190-229). If so, then the UIM would automatically include itself in the utilitarian equation. If its own consciousness exceeds that of all the rest of the world put together it might decide, on utilitarian principles, that men were redundant. If possible, I believe we should program or train the UIM so that it would not reach this conclusion. Personally I think it is unlikely that a machine could be conscious and feel pain because I think consciousness is embodied in the Schrödinger wave function (Good, 1962b, pp. 153, 335) or in something similar.

On the other hand I am not convinced that the route to the UIM would be via direct programming by humans. It seems at least as likely to me that it would be programmed by other machines; or that "built and trained" would be a better description than "programmed". There is no clear demarcation between programming and machine-building because of the possibilities of microprogramming. As a matter of history, my name for microprogramming was " " machine-building" " (in quotes) (Good, 1947): the name "microprogramming" was introduced by Wilkes (1951). But if the UIM is based on a simulation of the brain, and is therefore an ultraparallel machine, I don't think that either "programming" or "microprogramming" would be a good description because the complexity of the machine would be too great. That is why "training" might be a better description.

We should consider which would be better or worse: To have a UIM actually in charge, or to have it in an advisory capacity or in synergistic relationship with the boss. It might be less dangerous to have the UIM in charge because the machine might fall into the hands of unscrupulous men, and, as William Pitt, the Earl of Chatham, said in 1770 "Unlimited power is apt to corrupt the minds of those who possess it" (Evans, 1968, p. 547). Or, as Shakespeare (1601, I. iii), put it

"Then everything includes itself in power,
Power into will, will into appetite;
And appetite, an universal wolf,
So doubly seconded with will and power,
Must make perforce an universal prey.
And last eat up himself".

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.)

Anon (2058). *Handbook of Robotics* (56th edn.).

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.

Dennett, D.C. (1978). *Brainstorms* (Montgomery, Vermont: Bradford Books).

Evans, Bergen (1968). *Dictionary of Quotations* (New York: Avenel).

Good, I.J. (Feb./May, 1947). Unpublished notes on electronic computers, made in Manchester, England.

Good, I.J. (1962a). "The mind-body problem, or could an android feel pain?" in *Theories of the Mind* (ed. Jordan Scher; Illinois: Glencoe Free Press; second printing, with misprints corrected, 1966), pp. 490-518.

Good, I.J. (1962b). In *The Scientist Speculates* (ed. I.J. Good, A.J. Mayne, and J. Maynard Smith; London: Heinemann; New York: Basic Books; New York: Capricorn Books).

Good, I.J. (1965). "Speculations concerning the first ultra-intelligent machine", *Advances in Computers 6*, 31-88.

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).

Scriven, M. (1960). "The compleat robot: a prologomera to androidology", in *Dimensions of Mind* (ed. S. Hook; New York University), 118-142.

Shakespeare, W. (1601). *Troilus and Cressida*.

Thomas, Lewi~~n~~s (1980). "Notes of a biology watcher. On artificial intelligence", *New England J. Medicine* (Feb. 28), 506-507.

Wilkes, M.V. (1951). "The best way to design an automatic calculating machine", *Manchester University Computer: Inaugural Conference* (Manchester University), 16-18.

[Scanned 5/8/2018
by George Dyson
from a copy
kindly provided
by Jack Good,
October 1994]