precisely because I believe that the religion of the Incarnation-enfleshment must take account of gender, I believe that there is an urgent need for the Church to give much greater recognition both to women as people of the Church and to the significance of femininity in its worship and self-understanding.

1 My ideas about equality have been very much influenced by the French anthropologist Louis Dumont, who was stimulated by his encounter with the Hindu caste system to reflect at length on ideologies of equality and of hierarchy. See his *Essais sur l'individualisme,* Paris, Seuil, 1983.

2 See Ivan Illitch, *Gender,* London, Marion Boyars, 1983, for an ingeniously argued case that the incorporation of women in the labour market has not brought about their emancipation and that a society characterized by highly differentiated cultural and economic roles ascribed by gender would not necessarily be more unjust than present-day Western society, which professes to recognise simply the biological differences of the sexes.

3 An example of movements seeking for 'space' rathern than 'equality' would be contemporary North American Indian movements.

4 For a round-up of recent anthropological and theological views on sacrifice, see M. Fortes and M. Bourdillon (editors), *Sacrifice,* Cambridge University Press, 1980.

5 Islam, of course, is a fascinating example of a great religion which attaches very little importance to sacrifice, but sacrifice is significant in many forms of popular Islam.

The Theology of Robots

Edmund Furse

Artificial Intelligence: an introduction

The initial reaction of nearly all theologians and religious people to the very idea that it is possible to talk about 'the theological dimensions' of the existence of robots would—today—be dismissive, and, more often than not, scornful. 'It makes no sense,' most theologians would say. Before beginning to argue that one day, on the contrary, it will make a lot of sense, something much more general must be said about robots, or, more specifically, about Artificial Intelligence.

Artificial Intelligence, or AI as it is usually abbreviated, is the study of computer models of intelligent behaviour. Some scientists are interested in

https://doi.org/10.1111/j.1741-2005.1986.tb06559.x Published online by Cambridge University Press

using AI to understand human behaviour, others in designing intelligent mechanisms. As a discipline in its own right, it has existed since about 1950, with the pioneering work of John McArthy. It has two separate strands in its historical origin. Psychologists after the dark ages of behaviourism, which banished all talk of mental models as unscientific, started to study cognition (commonly called thought), which formed the subject of cognitive psychology. In devising models of mental processes they naturally turned to computers for an appropriate language of description; thus were born information processing models of human cognition. In order to formulate precise and testable theories of mental processes computer models were found to be indispensible. Models have been developed for memory, understanding natural language, vision, learning and, more recently, emotion.

Computer science has probably had a longer historical interest in AI, although one could argue that cogwheel and pneumatic models were an early attempt by psychologists to understand the mechanics of the mind. Many inventors have had an interest in devising intelligent machines; for example, machines which played games existed in the 19th century. There is thus a two-fold convergence on the study of intelligent behaviour. From below, by the gradual building of ever more sophisticated machines which exhibit intelligent behaviour; from above, by the devising of increasingly sophisticated models of human intelligence.

Some of the philosophical issues I will be turning to shortly are controversial and uncertain. What is certain, however, is that a great deal of money is being spent on Artificial Intelligence, and a great deal more will be spent in the future. The estimates for the world AI market in millions of US dollars for the years 1983, 1987 and 1990 are 100, 1500 and 4000 respectively. It is a truism that the motor car and the television have had a profound effect on our lives. Some of these changes may not have been desirable, but we would have had to think a long time in advance to prevent them. In contrast the effects of cars and television are quite insignificant compared to the potential effects of artificial intelligence on our society. Profound change may be many years away but I believe we need to examine the issues now before it is too late to alter the irreversible course of history.

AI Hypotheses: weak and strong

The belief that intelligent robots could be built in principle to behave like human beings is usually formulated more precisely in terms of the weak AI and strong AI hypotheses. The weak AI hypothesis states that it is possible to simulate human behaviour, whereas the strong AI hypothesis says it is possible to replicate human behaviour. An analogy is drawn by Aaron Sloman with a hurricane. In the weak hurricane hypothesis it is possible to simulate hurricanes on computers, but nothing actually gets wet. In the

strong hurricane hypothesis it is possible to replicate a hurricane and actually make things wet by artificial means.

The strong AI hypothesis, that it is possible in principle to replicate mental states in machines which behave much like you or me, has created a great deal of controversy, most recently being the subject of one of the Reith lectures of Professor John Searle. People can get very perturbed by the thought of the science fiction of robots becoming an actuality. I am going to examine arguments for and against this hypothesis, but I myself believe the strong AI hypothesis and the argument will be biased accordingly.

First, though, I think it would be helpful if we looked at our feelings about intelligent machines. Why are we uncomfortable about intelligent machines? Why do we want to make clear distinctions between human and machine intelligence? (Likewise, why do we want to distinguish between animals and humans? In this case it could be argued that we wish to justify our exploitation of animals.) There seems to be a fear of intelligent machines, but it is not clear on what this is based. Much of the presentation of robots in films is malevolent, and this may be the cause of our unease. Alternatively, we may feel that our place at the pinnacle of evolution and of God's creation is called into question. We may feel sibling rivalry as a species, fearing that our younger brother the robot will supplant us, not only in our jobs and endeavours, but in God's love. A third possible reason is that we may feel it is immoral to build a robot in the likeness of man, and hence of God. We feel uneasy about scientists conducting experiments on human embryos: tinkering with God's creation. There is a similar unease about building robots.

I will consider a machine intelligent if it behaves in an intelligent manner, that is to say if it exhibits intelligent behaviour. We do not class human beings as intelligent by looking inside their heads or dissecting them; rather, we deem them as intelligent if they behave intelligently. Likewise with machines; their intelligence should be judged by their behaviour. By defining intelligence in this way I am opposing John Searle, who consistently rejects this sort of approach, saying that it leads to behaviourism, which he says we know is false. I believe he is mistaken; it does not lead to behaviourism, and if you follow Searle's line you end up having to dissect a Martian before you can decide whether it is intelligent or not.

But of course you may object that this is not a helpful distinction since we now have to define intelligent behaviour. There are unfortunately two problems in this task. First, the attribution of intelligence to a behaviour can be transitory, predicated by understanding. To put it in simpler terms, if we see some people do a task, e.g. solve a maths problem, and we are very impressed, we may term this intelligent behaviour; if, however, they explain how they solved the problem, we may then classify it

https://doi.org/10.1111/j.1741-2005.1986.tb06559.x Published online by Cambridge University Press

as trivial—anyone could do it—and no longer consider the behaviour intelligent. Secondly, somebody may exhibit one behaviour which we classify as intelligent but several other behaviours which are far from intelligent. So we can have specific intelligent entities and general intelligent ones. A specific intelligent entity can perform one behaviour intelligently. Many computer programs (and some human beings, perhaps) come into this category, e.g. chess-playing programs, calculus-solving programs.

But could machines exist which exhibited many intelligent behaviours? I would wish to affirm that machines could exist which could do any intelligent behaviour that another human being could do, i.e. affirm the strong AI hypothesis.

The most important objection to strong AI is the suggestion that a mechanism cannot have free will. This has been expertly dealt with by Professor Margaret Boden in her book *Purposive Explanation in Psychology* and her other writings. My own argument will not do justice to her book and cannot necessarily be taken as equivalent. To say that one has free will is in computational terms to say that one makes decisions on the basis of one's own reasons, i.e. that for any particular decision one can list the possible choices, go through the process of thinking about the benefits of each of these choices, and choose one as a result of this thought. It does not matter if some finer grained analysis were to reveal that the reasons one produced could all have been predicted in advance. Thus we truly experience free will, even if at some more detailed level of analysis we are deluded. The same principles would apply to robots.

The usual objection to ascribing free will to robots follows the classical arguments about free will and determinism; namely that a mechanism can be completely and accurately described, hence its behaviour can be predicted in advance, and therefore it cannot exhibit free will. I believe this last 'therefore' is open to criticism. The question to ask is 'Who can predict the future behaviour?' If I can predict all my own future behaviour, then presumably I no longer have free will. But if my friend can predict all my future behaviour, but she does not tell me the predictions, then as far as I am concerned I still experience free choice in my actions. Even if my friend tells me that she knows my entire future life but will not tell me it, I in my everyday life will still make decisions believing them to be my own free choice. Thus a mechanism's future behaviour might be known to a human being or to another machine, but it itself may not be cognisant of its future behaviour.

There are, however, a number of other problems with this objection. *Firstly,* the complete and accurate description of the mechanism may not be possible. There already exist computer programs whose form evolves from interactions with several human beings, e.g. game-playing programs which learn by experience. It is conceivable that a computer program

might learn from its experience so fast that no human or group of humans would have sufficient time to comprehend its inner being. Although a print-out of its internal structure could be read, it could be so long that many groups of humans would have died before it had been read, and, anyway, it would have long since changed its inner structure by further processing. Secondly, even if a mechanism could be completely and accurately described, the prediction of its future behaviour might require too great a level of processing either for itself to perform or any other earthly computer. The simulation of logic circuits' behaviour for a few minutes of time can take many hours of computer time, even using large computers, if the circuits are sufficiently complex. It is not difficult to imagine a mechanism so complex that predicting its future behaviour in real time would require a computer larger than the known universe. Computers can run simulation programs which simulate the computer's behaviour. But such simulations take ten or more times longer to run than the events they describe. Thus a large mechanism may be able to predict ahead of the event the behaviour of a small mechanism, but a small mechanism could not predict its own future behaviour, and large mechanisms may not have their future behaviour predicted by anyone but God.

A *second* important objection to intelligent mechanisms is the belief that, however intelligent their behaviour might be, one cannot necessarily ascribe *consciousness* to them. If we are to avoid solipsism then we are forced to assume that the other human beings we meet are conscious of their own actions. Just so with machines. If we encounter a machine which communicates to us in terms of its own thoughts and intentions it is natural to ascribe consciousness to it. There does not appear to be any specifically human behaviour which (a) pre-eminently indicates a conscious mind behind it and (b) could not also be performed by a machine. Most peoples' experience of computer programs is of passive entities which obey the human user's commands. But there do exist programs already which take the initiative in dialogue, and in general express self-agency. Furthermore, some humans exhibit such simple behaviour patterns that their imitation by machines would be straightforward even with today's science.

A *third* objection to intelligent machines is the belief that, however cognisant they may be of their thoughts and observations of the world, they cannot truly *feel*—they cannot love or be angry. Not very much research has been done by AI researchers in studying emotion, but there are some workers in the field, e.g. Kiss and Sloman. In modelling emotion there are essentially two components: the first is an evaluation function which determines which emotion one is feeling and its intensity; the second a performance function which expresses the emotion through external action and may also effect internal processes. There is no reason in

https://doi.org/10.1111/j.1741-2005.1986.tb06559.x Published online by Cambridge University Press

principle why both of these functions should not be modelled by computer systems. Programs already exist which have internal monitors of their feeling states, e.g. Colby's PARRY, which simulates a paranoid.

The *fourth* objection to the strong AI thesis is John Searle's argument that programs are only syntactic objects and do not have any semantics. Whilst it is true that programs conform to a formal syntax, it is untrue to say that they do not have any semantics. Even the simple statement PRINT "HELLO" in the programming language BASIC has semantics, namely print the message HELLO.

I may have convinced you that generally intelligent machines exist in principle, but you may think that we will have to wait a long time for science and technology to produce them. This is a reasonable objection. The question is thus not 'Can there be intelligent machines?' but 'When?'. Nobody knows. At the 1983 International Joint Conference on AI, 75% of the delegates to a panel session on symbolic computation believed in the strong AI hypothesis, but no one was prepared to estimate when we would see intelligent machines. Sir Clive Sinclair stated at the 1986 European conference on AI that we would see androids by the year 2040, but most of the delegates thought this was too early an estimate. There is unlikely to be an overnight change, but a gradual increase in the intelligence of machines until they eventually surpass us. My own prediction is that we will see robots performing most of the intelligent tasks that we do definitely within 100 years, and possibly within 50 years.

It is important to remember that the argument for strong AI is an argument only *in principle*. Strong AI does not say that we can replicate minds using existing computer technology. Too much of the argument about strong AI is driven by peoples' images of existing computers rather than what is possible in principle.

AI: the implications for theology

This is not the place to explore the massive social implications of AI. We will focus on the implications for theology.

Already there is a growing exchange of concepts between AI and other specialisms. Some concepts from AI and computer science are already commonplace in psychology, at least in cognitive psychology, for example information processing models, stores, and programs. Concepts of goals, plans and agents are useful when talking about self-consciousness. Most existing programs are user-driven, but programs which take the initiative in dialogue and have their own goals have been developed. Intentions are of paramount importance when understanding human behaviour, and Daniel Dennett has argued that it is quite reasonable to use intentional language to describe the workings of complex computer programs. AI asks what goals does a person have, how are they related and ordered, and what plans are used to achieve the goals. Morals

can be seen as very high level goals.

Even today or in the near future various subdisciplines of AI have—or will have—contributions to make to theology. *Cognitive modelling* concerns itself with building computational models of mental processes, for example learning, memory, vision, and even neurosis and paranoia. In the field of human-computer interaction, models of the human user have been developed to distinguish one user from another, and various individuals have been modelled by computer, e.g. Barry Goldwater. A computational model of Jesus would be difficult to design since the gospel writers did not intend their work to be used for this purpose and clearly much needed data is not available. On the other hand, we have here a novel means of testing some of our theological claims, and computational models of certain individuals—for example, individuals who are very widely regarded as saintly people—could be instructive. Such a model could be tested by simulating real-life situations and seeing how it behaved.

Biblical exegesis has long drawn on ideas from other disciplines—for instance, sociology. *Natural language processing* gives biblical scholars a new tool to analyse the grammar of New Testament writers. *Computational models of learning* could be applied to improve catechetics, and prospective teachers of religion could use machines which learned to try their material on before introducing it to a human class.

These are examples of short-term applications. What, then, are the moral and theological implications of AI for humans and robots in the long term, if the strong AI hypothesis is true? Let us consider first what are the implications for our understanding of *homo sapiens*. In this very limited space it would be best to do this in question form:

Does AI help us to conceptualize more precisely our ideas about human development, and indeed what it means to be a human being? Can we assign personhood to agents whose behaviour we describe in intentional terms? Conversely, do we withdraw the attribute of personhood from agents whose behaviour we prefer to describe in mechanistic terms? If so, do we attribute person status to animals or to babies below the age of two? If an embryo or a foetus, given a sufficiently rich environment, is a potential human being, is not the same true of machines? Is the intention of the experimenter the crucial distinction? (In other words, if the experimenter is developing a computer program with the specific purpose of designing an intentional agent, can that program then be described as a potential intentional agent?)

It is clear that we are presupposing a purely reductionist model of human behaviour. What, you might ask, then happens to our theology? I believe that Christian theology is fully consistent with a mechanistic model of humanity. Earlier objections were based round the idea that mechanism entailed the denial of free will. AI has shown that to describe a being in a

mechanistic manner does not prevent one also using an intentional description. Mechanisms can be very complex, and if there is conscious free choice within the mechanism, then there is free will.

Robots' moral and theological problems

I intend to structure this closing section chronologically, dealing in turn with the 'birth', 'life' and 'death' of robots.

In the first place, who should create intelligent robots? If we compare the creation of robots to human birth, how are we to ensure adequate moral practice? Should each new creation of a robot be registered? Should individuals or institutions need licences to create robots? Should robots be allowed to create copies of themselves? None of these questions are easy to answer and combine moral and legal issues. Since intelligence is not likely to be suddenly acquired but rather is a continuum, it is not necessarily appropriate for the discussion to be purely in terms of switching on an intelligent robot. Creating a robot can be compared to creating a child, but, just as children can be brought up in different ways (e.g. in a family, or in a kibbutz), so also there are likely to be many ways that a robot can be brought into existence and learn about the world.

Moving on to consider the 'working life' of robots, many of the moral issues have already been examined in science fiction, and the issues of robots' rights have been rightly identified by Asimov. Truly intelligent robots will not be totally subservient to humans, but will think for themselves. It may be possible, and indeed desirable, to incorporate Asimov's three laws into all robots:

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.

But this is not easy to integrate with the robot being a creative autonomous agent. What do you do when you say to your robot 'Would you do the washing-up, please?' and you get the reply 'I would like to finish reading the paper first'?

Perhaps the most fundamental question to ask is 'Would robots sin?'. This is a very interesting question since if there is a possibility of answering 'No' it has important implications. Of course, it would not strictly be correct to say that any subject (be it human or robot) 'sinned', even if it behaved in an 'immoral' way, unless its existence was orientated to a transcendent goal, at least potentially, and it sensed that. The question whether or not a robot could in any way be said to have a 'religious' life we will consider shortly. First, however, we must consider something easier to grapple with. A subject (in this case a robot) cannot be said to 'sin' unless

it knows what it is doing, is exercising free choice, and knows that the action it has chosen is morally wrong.

In order for the robot to know that an action is morally wrong it would need at the very least to have some moral code built in. Kant would argue that it could not deduce moral principles from factual knowledge of the world. It could determine what the social consequences of an action would be and, by a process of deduction from the built-in norms, deduce not only whether an action was undesirable but whether it was immoral. The robot will have certain goals it is following at a given time and a number of different means for achieving them; it may choose the means on a basis of efficiency or other criteria, and these criteria could include moral principles.

A more difficult question to answer is: Is it imaginable that the robot should want to do something wrong, and if so, why? Presumably it could create a goal it could only achieve by immoral action, but how? Obviously a human could give it an immoral goal, but in that case we could hardly say that the robot had 'acted immorally', and certainly not say that it had 'sinned'. A more fundamental question is whether the robot could acquire an immoral goal without it being explicitly added, and it is surely conceivable that since the robot exists in a corrupt and complex world it may inevitably acquire immoral goals; it may be forced to choose between the lesser of two evils.

The Christian will presumably argue that it must be possible in principle for a human to behave in a perfectly moral manner, in spite of being fully aware of possible immoral goods. The AI researcher must ask how can this be achieved; what are the processing implications of a perfect moral life? When new goals are acquired they will need to be analysed for their moral implications. But this could be an extremely complex undertaking.

The reasons for a robot's actions could in principle be available for inspection by a robot, other robots or humans. Many expert systems today allow for a human to ask the program to explain its decisions. If robots could always be asked to explain their actions, to any desired level of goal structure, then they would be more trustworthy.

If a robot is curious (which it will probably have to be to function effectively) and if it takes an interest in the world (by reading, absorbing television material, or other means, which it will need to do in order to relate to human beings), then sooner or later it will stumble across words such as religion, God and sin. If a robot asks you if you could find it a copy of the bible to read, what do you do? Surely you would get it a copy. But would the robot understand the bible? If the strong AI hypothesis is true, then presumably the answer is yes, at least at a cognitive level. Since we argue that it is a very reasonable thing to be a Christian, surely those of us who are Christians would expect that a robot, the most reasonable of

https://doi.org/10.1111/j.1741-2005.1986.tb06559.x Published online by Cambridge University Press

sapient beings, would want to become a Christian. But we also say, of course, that this requires the grace of God—and it is not obvious that God desires the salvation of robots as well as homo sapiens.

Is it, however, preposterous to think that God might desire the salvation of robots? Arguably there are beginnings of a foundation for the claim that robots could be 'called to God' in Karl Rahner's theory of the resurrection as a new relationship with the world; St Thomas Aquinas argued that all creation is orientated to its ultimate goal, and (quoting Rahner again) it is through man that nature is first touched by God's grace. It is not only human beings who are redeemed, but—ultimately—all creation. And surely the most remarkable non-human thing that human beings are giving birth to, the intelligent reflective robot, has a very high place in that redemptive order?

Finally, a word about the robot's death. Although switching off a robot can be compared to murdering a human being, the comparison is not straightforward. Since robots could presumably live much longer than humans, indeed in principle arbitrarily longlife times, it might be desirable for them to die. Putting this another way, would humans be prepared to share the planet with a race of immortal robots? We are more likely to have to face the immortality issue with robots than with humans. Some researchers have even suggested that humans could transfer their knowledge to robots, thus living out their existence in a non-biological form. But although this may be possible in principle, it will be much more difficult to achieve than autonomous robots.

What then, is our conclusion? Strictly, the avoidance of a conclusion. Artificial Intelligence can be seen as the latest progression in the use of computers in society. But AI also involves qualitative change over previous computer programs, and I believe has profound implications for life on earth. As a Christian and an active AI researcher I think it is necessary to bring some of the issues forward for discussion, and particularly some of the moral and religious issues, so that we can decide what sort of world we want to live in—assuming, of course, that the nuclear holocaust does not overtake us first. This is not the end of the discussion, but only a very modest beginning.

BIBLIOGRAPHY

Boden, Margaret A., *Artificial Intelligence and Natural Man,* Harvester Press 1977.

Boden, Margaret A., *Purposive Explanation in Psychology,* Harvester Press 1978.

Boden, Margaret A., *Minds and Mechanisms: Philosophical Psychology and computational Models,* Harvester Press 1981

Dennett, Daniel C., *Brainstorms: Philosophical essays on Mind and Psychology,* Harvester Press 1981.

Searle, John, *Minds, Brains and Science: the 1984 Reith Lectures,* BBC Publications 1984.