# 1

# Social Implications of Intelligent Machines

R. L. Gregory
Brain and Perception Laboratory
University of Bristol

This paper is designed to have a high coefficient of fiction. This should not be taken to mean that it will necessarily be false; but rather that it will play with possible realities. It will indeed be argued that playing with possible realities is the essence of intelligence; and that machines will not be intelligent until they are designed to function not by direct control from events, but rather from a continual running internal fiction of the world of events.

What do we mean by intelligence? There is no agreed definition; and psychologists are apt to confuse processes leading to intelligent solutions with what it is to say that a solution is intelligent. But it is confusing to equate, say, good memory or concentration, or any such, with intelligence – even if these characteristics are necessary for deriving intelligent solutions. It is confusing because we should be clear about what it is to say that one *solution* is intelligent, another not, irrespective of how the solution is obtained. To judge that a solution is intelligent we do, however, have to know what data and what previous solutions were available, or we will be in danger of attributing intelligence to something that merely copies. Clearly a necessary criterion of intelligence is novelty. Novelty alone, however, is not enough, for what is novel may be arbitrary, or downright misleading. Evidently, to be appropriate is also a necessary condition for intelligence. I shall suppose that these two criteria are sufficient for defining intelligence. I shall proceed to define an intelligent act, or an intelligent solution as *any act, or solution, which has appropriate novelty.* This definition should allow intelligence to be quantified, and measured.

Whether it is a man or a machine which is responsible for an example of appropriate novelty, that man or that machine will be called intelligent. It is perhaps not quite clear how far appropriateness can be quantified; but certainly novelty can be quantified – in terms of prior probability. That man or that machine which succeeds in producing appropriate solutions

having the lowest prior probabilities will be declared the most intelligent man, or the most intelligent machine. Others will be graded according to the lesser novelties of their solutions; while any 'solutions' which are not appropriate will be ruled as non-intelligent. Appropriate novel solutions might, of course, occur by chance, but this will be exceedingly unlikely to happen for a given man or a given machine on many occasions; and so if it does occur on many occasions we may be sure that that particular man or particular machine is indeed intelligent – and so is likely to produce further appropriate novel solutions.

We may now ask: How did intelligence develop in organic evolution? It is clear that early organisms do not show intelligence, according to our definition, for though their actions are often appropriate, they are seldom novel. Behaviour of simple life forms is essentially reflex, actions being initiated by rather specific stimuli. Whether a given reflex is exhibited may depend on the hormonal or other state of the organism, and what is elicited may be a complex series of actions, forming behaviour patterns; but reflex behaviour can be described fairly adequately in terms of the setting of conditional circuits – so that an input directly triggers an output response. This may be appropriate but will not be novel – and so it will not be intelligent.

Stimulus response, or other direct-control-of-output-by-input systems, has essential limitations. It is in terms of overcoming these limitations that we see the development of intelligence in evolution. The first limitation of systems controlled directly by inputs is that they are lost when their inputs fail. Mechanical systems, such as cars, are lost when their control wheels come off, and a sophisticated servo-follower is lost when it loses its information link with its target. In general, machines stop, or their output becomes inappropriate, when their inputs fail: but this is not true of the higher organisms in spite of the fact that the problem is acute for organisms because the flow of sensory data is extraordinarily unreliable. Also, very often what is available is strictly inappropriate to guide the task in hand. Organisms keep going through gaps in the flow of sensory data quite remarkably well. They also succeed in behaving appropriately to characteristics of objects which are not monitored by their senses. For example, we generally pick up the cool end of a soldering iron, without having to monitor the heat we avoid. Now what does this imply? The ability to behave appropriately during data-gaps implies quite directly that we are not merely reflex-response systems, as some psychologists have supposed. Since we are able to bridge data-gaps by effective assumptions of what is going on, it follows that *behaviour is controlled by assumptions of the state of the world.* I shall call such assumptions of external states *fictions.* When used to predict future states they might be described as *hypotheses.* The brain's fictions may closely correspond to aspects of reality: the term 'fiction' should not be taken to imply that they are false – any more than all literary fiction is quite false. Just as a story is based on past experience, and may correspond with the present or the future, so the brain's fiction may be appropriate and so useful.

The special power of brain fiction is that it frees behaviour from the tyranny of immediate sensory control. It seems reasonable to guess that it first developed to bridge gaps in sensory data, and that the first brain fiction was no more than simple extrapolations of observed trends: to bridge the unsensed present by projecting the past into the future. This does require the tacit assumption that nothing drastic will happen during the gap, and it is bound to fail if conditions change too much. A failed prediction of this kind may be novel but it will not be appropriate, and so it should not be regarded as intelligent. Data-gap filling, although useful and a necessary step toward intelligence, is not itself intelligent because it is not able to generate appropriate novelty.

We may describe data-gap bridging as *cognitive inertia.* It is important not only for what it led to in evolution, but also for greatly increasing performance reliability when the available input of data is intermittent, as it generally is for organisms.

We have already hinted at what seems to have been the second important step in the evolution of intelligence: the ability to read hidden features of the world from what is given by the senses. One example is avoiding the hot end of the soldering iron; another is accepting an ice cream, on the evidence of the retinal image which itself is not cold, heavy, sweet, or edible. These characteristics are read from the retinal image; much as we read from a book, say, that a lighthouse stands on a cliff. We should say that the image of the ice cream has selected a *fictional account*, stored in the brain, of ice creams and what they can do, and what the brain's owner can do to them. Behaviour is but distantly controlled by the retinal image: it is controlled by the brain's fictional account of ice creams, cliffs, and lighthouses – by a host of objects and situations from the past. Brains then have the possibility of generating appropriate novelty; for they have but to present items of stored fiction to each other, or to sensed situations, and they may discover appropriately novel solutions. The discovery may appear, from the outside, as a unique creation.

To produce intelligent machines we might repeat this supposed development of organisms – to produce machines controlled not by direct information from the world but rather by their own fiction. We can however hardly expect that the machines' fictions will be like ours: but this is only a small part of the problem of predicting their effect on human society. First we should consider the social effects of *non*-intelligent machines; particularly whether we can predict *their* effects.

## SOCIAL EFFECTS OF NON-INTELLIGENT MACHINES

Up to now almost all machines have been passive slaves, controlled by inputs from the world or by human commands. A train is guided directly by its track, and power tools cut according to instructions. Instruments, such as rulers and sextants, give readings directly of selected features of the world. There is however one machine, invented in its modern form six hundred

years ago, which is fundamentally different. This machine is not under continuous, but only very occasional, control or interference from outside. It has a kind of inner secret life and the answers it gives are, when appropriate, most useful. I refer to the clock. Clocks do not record time directly. Indeed, we do not know what it would be to record time directly. A clock works by 'living' an inner fictional time; which we can read from the gestures of its hands on its inter-face.

A clock is the extreme case of a system which is useful because of its inertia. By plodding on regardless, and not responding to particular events, it can mirror the average change of things and in this it is useful. If a clock departs from what it is set to represent it becomes inappropriate, misleading without striking analogies. We should say that a clock is never intelligent: for just when it is appropriate it is not novel; and when it is novel then it is no longer appropriate.

Clocks show us that inertia, though useful for filling data-gaps with fiction which might be appropriate and useful, is nevertheless not adequate for intelligent machinery. We may however say that the early clock-makers paralleled the first crucial step in organic evolution towards intelligence – they made the first machines to work by fiction.

Clocks have had quite large social effects, but not especially because they function by inertia. They do strike us with a special awe, as subtly different from other machines, but there are many machines which have had more effect on human life. The inventions of the plough, the lathe, and a thousand others must have had more social effect than clocks. Most inventions are to some extent anticipated and few have the startling implications of machine intelligence yet all important inventions seem to catch us unawares, to produce unpredicted results. Perhaps it is science fiction writers who have the best record for accurate prediction; but they are more often wrong than right. Jules Verne was the most accurate in his predictions but even he saw no engines beyond beam engines. His future floating island, for example, had no radio but had to plug into undersea telegraph lines existing in his time. Space travel seemed impossible to professional astronomers before the 1939 war; and the Chief Engineer of the BBC declared firmly that television was theoretically impossible, at about the time it was demonstrated by John Logie Baird.

The history of inventions sometimes makes one wonder how far we are intelligent, and how far we are merely inertial even at peaks of imaginative creation. An excellent source of cases is Samuel Smiles' *Lives of the Engineers* (1862), which gives fascinating details of the difficulties and hang-ups of inventors. As an example of what appears to be inertial thinking, we may take the first stages of the design of railway engines. Several inventors had difficulty in imagining that it would be possible to apply power to wheels to make a vehicle move. Even Trevethick, in his master patent of 1813, stated that: 'the driving wheels should be made rough, by the projection of bolts or

cross-grooves, so that adhesion to the road might be secured'. In the same year a Mr Brunton, of the Butterley Works, Derbyshire, patented his Mechanical Traveller, 'to go *upon legs*, working alternatively like a horse'. Unfortunately the boiler burst on its first and only trial. In 1814 Thomas Tindall designed a locomotive in which, 'the power of the engine is to be assisted by a *horizontal windmill*; and the four pushers, or legs, are to be caused to come successively in contact with the ground, and impel the carriage'. The point is that wheels on vehicles up to that time had been passive, as carts and carriages were pulled by animals. The notion of powered wheels was novel and evidently difficult to grasp, even after it had been suggested, as it was not part of the brain fiction of the time.

To us, looking back at this, steam horses look like a clear symptom of cognitive inertia rather than intelligence. A rather different example of cognitive inertia is the transfer of ideas from bird to human flight. At first the dynamic wings of birds were copied by inventors, leading to several flapping human deaths.

From the social history of inventions it is clear that important effects arise from completely unnoticed origins. An interesting example is how the use and limitation of horse-drawn buses and trams led rather directly to the building of houses of poorer people around the growing towns in valleys, rather than on hills, though the valleys were difficult to drain and so were unhealthy for large numbers. The hills remained the preserve of the rich largely because public vehicles, with their heavy loads, over-taxed the strength of horses, while privately-owned vehicles carrying only a few passengers were not too heavy. This situation continued until engined buses could manage hills (when in any case the drainage problem was solved) and now we can all go to work from the heights. This was not predicted or planned – it just happened that way. But what of intelligent machines? If the social effects of horse buses and petrol engines cannot be anticipated, what hope have we for the onset of intelligent machines? How can we hope to overcome our cognitive inertia – to use our intelligence to guess correctly for such a novelty?

## SOCIAL EFFECTS OF INTELLIGENT MACHINES

The only hope we seem to have for predicting the effects of intelligent machines is that intelligence already exists. Let us start by assuming that our intelligent machines will be metal men with similar intelligence and with inner fictions that are similar to ours. Further, we will suppose that in their construction they will be typical electro-mechanical machines.

Such machines would have, at the very least, the advantages of brave men dressed in asbestos armour; eating little or nothing, and perhaps with a tranquillizer so effective that they can be made to sleep for centuries, to be awakened when they reach distant planets, or to teach our descendants intimate history. These metal men would be of great help in war and in all kinds of dangerous and boring industrial jobs. It seems all too clear however,

from our reaction to people with different origins and only slightly different looks, that even if the brains of the metal men were essentially like ours we would hardly accept them into the human club. At best, they would be another and very odd race, which we would have every excuse to exclude from human ethical restraints – as we would have no reason to believe them capable of experiencing pain. This is, of course, a science fiction theme but it should not be quite dismissed on that account. Science fiction can be regarded as the first probings of the imagination; the first attempt to consider a problem if not solve it. Indeed, the ancient cosmologies were just this. We no longer believe that the earth stands on an elephant, but the very asking of the question led to appropriate observations and formulated theories giving precise predictions. The important point is that neither primitive cosmologies nor science fiction are safe guides to prediction. This is clear from their contradictory variety. We could write a plausible science fiction story in which the metal men are accepted by human society, but with the result that men lose their confidence because the flesh is so obviously weaker than metal. We could write another story, in which the women become so fixated upon the metal men that the human race dies out for lack of gene pairings – or even that the genes get polluted with iron filings! In still another story we might invent, human work and decisions are taken over until men give up all serious things to play only games. The final game is the destruction of the metal men – leading to the end of the human race, as men have forgotten to live at all like animals. Now the point of these stories is that they are all possible fictions. Although referring to possible futures, they are however no more than games with familiar ideas as counters. But can the future be described at all adequately in terms of present-day situations? The essential limitation of this kind of primitive – inertial – prediction is that essentially novel possibilities are excluded. It is therefore non-intelligent. It is just this which is the weakness of history used for prediction.

Surely we cannot predict the effect of intelligent machines on this basis. Is it possible to base predictions on anything more likely to succeed than stories limited by cognitive inertia? Surely physical science makes at least some kinds of prediction possible: can the methods of science help us to predict the social effects of intelligent machines?

To take an early example of prediction by the methods of the physical sciences, we may consider the prediction of eclipses. Solar eclipses have been predicted, with fair accuracy, for perhaps four thousand years. Presumably it was noted that eclipses occur only at full moon, and under certain other conditions occurring in cycles, allowing prediction once these cycles had been mastered. In fact the heuristic program required for prediction of *all* eclipses at any place on earth is extremely complicated and complete accuracy was not attained until a conceptual model of the solar system was developed. Prediction was then not in terms of the solar system as observed, but in terms of the conceptual fictional model. It was only when the observations became

secondary to the model that prediction became complete. Further, it was possible to predict not only eclipses but a host of other phenomena. It became also possible to introduce quite new factors, space ships, into the situation – and, finally, to touch the moon.

It is particularly interesting that as prediction became more reliable, and the planets were seen to move along defined orbits, the notion that they are intelligent, or are pushed along by intelligent beings, was dropped. This is compatible with our definition of intelligence for as we develop the power to predict so there is less novelty in the events predicted. Scientific theories destroy the appearance of intelligence in things, as prediction becomes possible. This is true for biology as it is for physics. This does however lead to a paradox for sociology.

Sociologists are concerned to predict the effect of changes on future society. But is prediction *in principle* possible when intelligence is involved? If intelligence is the production of novelty, accurate prediction might seem to be strictly impossible. However this may be, it seems that the present trouble about social prediction is simply that there are no adequate theoretical models of societies. This means that politicians are almost powerless to predict, plan, or control, except with incredible errors. We find ourselves in just this position in trying to assess the implications of future intelligence. We are in the position of the early astronomers with no model of the solar system.

In these circumstances the best we can do is to write fiction from our past, and just hope that the story we like best turns out to be true. Without a theoretical model of society we can do no more than adopt inertial procedures, and accept that our predictions are virtually certain to be neither appropriate nor novel.

The vital point about intelligent machines is that once they are trusted they will take decisions, and these decisions will affect us directly. In a sense present-day technology makes decisions – certainly it sets up situations beyond our power to predict or prevent. Intelligent machines will be a fundamentally different case when we ask for their opinions. Perhaps the most important questions here concern the ethics of responsibility. Suppose a mechanized judge (programmed with the law, and the relevant data of the case) condemns a man to punishment. Now suppose that, after the punishment is inflicted, the mechanical judge is found to be in error – what would our attitude be? We might assume that the machine went wrong – that some electronic malfunction was responsible for the error. We might then blame the designer or the maintenance staff – at least if they were human – but surely not the judge-machine, any more than we blame our car when the battery is flat. Suppose that it was clear that the machine's components did not fail – but rather that the machine had to balance probabilities and on this occasion the most likely candidate happened to be innocent. This is bound to happen for human judges and it is also bound to happen for machine

judges, for it is not always the most probable which occurs. Indeed if it did there would be no useful concept of probability. Would we blame the machine when it is misled by the improbable event happening to occur? We might – rationally we should not. We might, because we do tend to blame people when this happens, in a serious situation, such as a court of law. But is our blame rational? I think it is not.

Of course to blame or censure a judge for an error may serve to sharpen his future judgement, and no doubt a similar procedure could be appropriate for machine judges. This would however be called optimizing the machine's procedures, rather than punishing or censuring it. Possibly, as this becomes the general practice for dealing with machines which come up with incorrect answers, our notion of punishment and of blame may change – and in becoming adapted to the machine we may ourselves become more humane.

If machines are to make decisions on human affairs, then certainly they will have to be programmed (or will have to discover for themselves) a great deal about the behaviour, aspirations, and fears of human beings. More profound: in order to design human-like intelligent machines we will have to make psychology a far more effective science. There are indeed already signs that the study of machine intelligence is affecting experimental psychology. Here we come back to the inner fiction, the brain's symbolic models, describing features of the external world. It is surely vitally important to go beyond stimulus-response psychology, and to accept clearly that the prevailing sensory input is but a small part of what determines human behaviour. It is surely the detailed studies of cognitive structures which will be the effective description of man. It is these structures, no doubt partly reflected in the structure of language, which will be the essential design descriptions for the intelligent machines.

## PERSONAL RELATIONS WITH INTELLIGENT MACHINES

We may be correct in concluding that it is impossible to predict the effects of the introduction of a new feature into society, but we should still ask: Is it possible to predict the 'psychological' relation between people and intelligent machines?

Unfortunately, we seem to be forced to the view that just as an adequate theoretical model is necessary for predicting effects of a novel feature into society, so a detailed model is required to predict individual reactions. The fact is that we know so little about people that historians show almost no agreement even over which past events should be regarded as causally related to later events, or to later individual or social attitudes. If the past cannot be interpreted in this way – when the range of possibilities is limited by what is known to have happened – what hope have we to predict the future, when the possibilities are limited only by logical considerations? Unfortunately we must adopt the inertial procedure, arguing on the basis that we do know what it is like to deal with intelligent organisms, especially other people having

roughly similar fictions, hoping the case of intelligent machines is not strictly novel. There are certainly some considerations which apply both to other people and to intelligent machines of human capability.

This brings us to our final consideration, which is: What happens when the internal fiction of the machine is very different from human brain-fiction? This will surely be the case, for even if we understood human brain-fiction in detail, which certainly we do not at present, it seems most unlikely that the kind of software developed through evolution for survival in past conditions would be optimal for machines designed specifically to solve problems – even if they are *our* problems. Human emotion may be important for selecting immediate priorities, in terms mainly of survival and reproduction, but the selection of data and aims set by emotional states would surely be inappropriate for machines having very different survival and reproduction problems. Human prejudice is useful in saving thinking time: clearly it would be intolerable to have to consider all relevant possibilities. No doubt pre-selections of possibilities, which we might as well call 'prejudice', would have to be accepted by the machines also, but there seems no reason why their prejudices should agree in any detail with ours. The trouble is that if they do not, communication is certain to be extremely difficult. Just as it is very difficult to communicate across prejudice (or opinion) barriers between people, so it will be equally difficult or impossible with machines. The power of philosophical discussion is, surely, to make explicit underlying assumptions in human arguments: if the software of machine arguments is totally exposed, and the machines are pitted against us in debate, then we can expect exciting and perhaps too challenging clarification of human thinking. This will be an extension of the effect that computers are already having on our understanding of logic and mathematical procedures. The hope is that intelligent machines may reveal where we are arbitrary and non-rigorous in our use of language having semantic content, much as existing computers show up inconsistencies in formal orderings of symbols apart from what meaning we may attach to the symbols.

Apart from the sheer difficulty of reproducing human brain-fiction it is most unlikely that it would be worth while; except that it may be essential if we are to communicate directly with the machines. One can imagine a class of machines which works quite mysteriously, with non-human fiction, to give us answers without justifications we could understand. Some people might come to trust such machines, much as they trust cars though they have no idea how the steering wheel is connected to the front wheels. But would it be possible to phrase questions appropriately to such machines? It seems more likely that these alien machines would be outside direct human control, but would feed themselves with raw data through their own sensory systems, and be left to find answers to problems we may but vaguely understand. These machines would form a separate race of hidden intelligence; which could come up with devastating novelty which we might be hard pressed to

find appropriate. We might find it difficult to accept the decisions of very intelligent machines as appropriate, and so as intelligent.

There is a related point here of importance: an issue which is already with us, as computers are used to store and handle personal and confidential details of individuals. Quite apart from the intelligence of the computers, or lack of it, we tend to feel threatened by this impersonally-stored mass of data, which can be retrieved at a moment's notice by anonymous bureaucrats. If the bureaucrats were machines, we might feel much the same. The worry is the threat to personal liberty implied by the ready availability of facts about ourselves. On the other hand, it might be admitted that administrative decisions would be more effective, and perhaps more just, if adequate facts were available. There is a conflict of opinion here, though no logical conflict for both opinions could be right. We may expect a loss of individual freedom, at least to get away with minor transgressions, while perhaps attaining more rational government.

How far personal information should be collected and made available for computer handling can only be judged in terms of a preferred fiction of the future – what sort of world we wish to leave to our descendants. Here we reach a curious ethical question.

It is clear that people living at various times and in different societies have somewhat different moral standards and preferences. Further, people generally accept the moral standards and preferences of their own society; at any rate if these do not change too rapidly. Since we tend on the whole to accept our own society as we find it, what right have we to inhibit present developments, on the grounds that our values will continue to be held by later generations? In fact, we may assume that changes made now will become more acceptable as time goes on. Indeed value judgements may represent no more than cases of extreme, though not complete, moral inertia. To take an example of this kind of situation, consider the Victorian attitude to the telephone. For many years it was regarded as an intrusion of the privacy of the home, for it was every middle class man's right to be 'not at home' to visitors, and this principle was violated by the telephone. The Victorians might have banned the development of telephone exchanges handling private numbers, and so inhibited at least for several years the general use of telephones. Now although this would have seemed a good idea to them, does it now seem a good idea to us? To generalize this question: Are there technical developments which were foreseen and which we now wish had been suppressed for social reasons? There may be a few, but generally we seem to adapt to and accept the results of technical innovation. At present we do not in any case know how to suppress technical innovations; and so we have to hope that people will adapt to their social effects, or will invent adequate counter measures. This has worked quite well for the results of artificial power and non-intelligent machines. We can only hope that the introduction of intelligence to machines will not result in situations beyond

human ability to accept or to counteract. At least we should be able to harness the power of the intelligent machines to tackle their threat to humanity, as now we use machines to counteract ill-effects of other machines.

Although ethical notions change with time and across societies, no doubt some notions and restraints are shared by all societies. These could be programmed into the intelligent machines and in this way they could be made to 'see' our morality. But as human morality changes, should the morality of the machines be made to follow? More generally, should we build human ethical and other inertias into intelligent machines? It would certainly be intolerable for us to be judged by values accepted in past times; so presumably we should wish the machines to follow our social mores as they change, rather than (like ideal clocks) be perfectly inertial, in the hope of representing ultimate values.

Turning from the 'seeing' of situations and events in terms of ethical values; there are deep problems over what it would be to say that machines 'see' the common physical objects we take for granted. Here is the most difficult question of all: In what sense can a machine share our world?

This question might be phrased in our terms as: How far can machines share our fiction? It has been argued in other places by the author that perceiving the world involves a kind of *non-formal* intelligence; an intelligence not dependent upon language, but perhaps providing the origin of language. It seems likely that perceptual intelligence has to be non-rigorous, not strictly analytical, in order to come up with useful solutions – perceptions – in real-time. We can expect the same from seeing machines; but the assumptions they will adopt, to make the problems of perceiving objects from images tractable, will no doubt be different from our simplifying assumptions. If so they will, in an important sense, see a different world. We will be able to understand the machine's behaviour to its world only as we partly understand the behaviour of animals different from ourselves. It may be difficult to work in close cooperation with such machines. Again, we seem to see intelligent machines as having impenetrable fictions, engaged in rather mysterious ways even on the tasks which we set for them. In the long run this may be to our advantage, for it will remain true that we are special and so not directly challenged by the machines. No athlete is worried by the fact that horses run faster than any man, and mathematicians are not bothered by the computer's superior ability at arithmetic, presumably because horses, and computers, are sufficiently different from us. On this basis a world with intelligent machines could be not only interesting but compatible with human happiness; providing the machines are very different from us – or carefully programmed to show due tact to their masters.