REPRINTED FROM

ADVANCES IN COMPUTERS, VOLUME 6

ACADEMIC PRESS INC., NEW YORK

(397)

Misprints
Gme
42 (3)
54

Speculations Concerning the First Ultraintelligent Machine*

IRVING JOHN GOOD

Trinity College, Oxford, England and
Atlas Computer Laboratory, Chilton, Berkshire, England

1. Introduction

The survival of man depends on the early construction of an ultraintelligent machine.

In order to design an ultraintelligent machine we need to understand more about the human brain or human thought or both. In the following pages an attempt is made to take more of the magic out of the brain by means of a "subassembly" theory, which is a modification of Hebb's famous speculative cell-assembly theory. My belief is that the first ultraintelligent machine is most likely to incorporate vast artificial neural circuitry, and that its behavior will be partly explicable in terms of the subassembly theory. Later machines will all be designed by ultra-

* Based on talks given in a Conference on the Conceptual Aspects of Biocommunications, Neuropsychiatric Institute, University of California, Los Angeles, October 1962; and in the Artificial Intelligence Sessions of the Winter General Meetings of the IEEE, January 1963 [*1*, *46*].

The first draft of this monograph was completed in April 1963, and the present slightly amended version in May 1964.

I am much indebted to Mrs. Euthie Anthony of IDA for the arduous task of typing.

intelligent machines, and who am I to guess what principles they will devise? But probably Man will construct the *deus ex machina* in his own image.

The subassembly theory sheds light on the physical embodiment of memory and meaning, and there can be little doubt that both will need embodiment in an ultraintelligent machine. Even for the brain, we shall argue that physical embodiment of meaning must have originated for reasons of economy, at least if the metaphysical reasons can be ignored. Economy is important in any engineering venture, but especially so when the price is exceedingly high, as it most likely will be for the first ultraintelligent machine. Hence semantics is relevant to the design of such a machine. Yet a detailed knowledge of semantics might not be required, since the artificial neural network will largely take care of it, provided that the parameters are correctly chosen, and provided that the network is adequately integrated with its sensorium and motorium (input and output). For, if these conditions are met, the machine will be able to learn from experience, by means of positive and negative reinforcement, and the instruction of the machine will resemble that of a child. Hence it will be useful if the instructor knows something about semantics, but not necessarily more useful than for the instructor of a child. The correct choice of the parameters, and even of the design philosophy, will depend on the usual scientific method of successive approximation, using speculation, theory, and experiment. The percentage of speculation needs to be highest in the early stages of any endeavor. Therefore no apology is offered for the speculative nature of the present work. For we are certainly still in the early stages in the design of an ultraintelligent machine.

In order that the arguments should be reasonably self-contained, it is necessary to discuss a variety of topics. We shall define an ultraintelligent machine, and, since its cost will be very large, briefly consider its potential value. We say something about the physical embodiment of a word or statement, and defend the idea that the function of meaning is economy by describing it as a process of "regeneration." In order to explain what this means, we devote a few pages to the nature of communication. (The brain is of course a complex communication and control system.) We shall need to discuss the process of recall, partly because its understanding is very closely related to the understanding of understanding. The process of recall in its turn is a special case of statistical information retrieval. This subject will be discussed in Section 5. One of the main difficulties in this subject is how to estimate the probabilities of events that have never occurred. That such probabilities are relevant to intelligence is to be expected, since intelligence is sometimes defined as the ability to adapt to new circumstances.

The difficulty of estimating probabilities is sometimes overlooked in the literature of artificial intelligence, but this article would be too long if the subject were surveyed here. A separate monograph has been written on this subject [*48*].

Some of the ideas of Section 5 are adapted, in Section 6, to the problem of recall, which is discussed and to some extent explained in terms of the subassembly theory.

The paper concludes with some brief suggestions concerning the physical representation of "meaning."

This paper will, as we said, be speculative: no blueprint will be suggested for the construction of an ultraintelligent machine, and there will be no reference to transisitors, diodes, and cryogenics. (Note, however, that cryogenics have the important merit of low power consumption. This feature will be valuable in an ultraintelligent machine.) One of our aims is to pinpoint some of the difficulties. The machine will not be on the drawing board until many people have talked big, and others have built small, conceivably using deoxyribonucleic acid (DNA).

Throughout the paper there are suggestions for new research. Some further summarizing remarks are to be found in the Conclusions.

2. Ultraintelligent Machines and Their Value

Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind (see for example refs. [*22*], [*34*], [*44*]). Thus the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.

In one science fiction story a machine refused to design a better one since it did not wish to be put out of a job. This would not be an insuperable difficulty, even if machines can be egotistical, since the machine could gradually improve *itself* out of all recognition, by acquiring new equipment.

B. V. Bowden stated on British television (August 1962) that there is no point in building a machine with the intelligence of a man, since it is easier to construct human brains by the usual method. A similar point was made by a speaker during the meetings reported in a recent IEEE