Irving John Good - Speculations concerning the first ultraintelligent machine (1966)

Tom Rochette <tom.rochette@coreteks.org>

January 17, 2018 — 8330fe3

0.1 Context

0.2 Learned in this study

0.3 Things to explore

- An ultraintelligent machine can rely on other machines, like humans relies on machines to do things (as well as other human beings)

1 Overview

2 Notes

2.1 1. Introduction

- 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

2.2 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
- Thus, the first ultraintelligent machine is the last invention that man need ever make, provided the machine is docile enough to tell us how to keep it under control
- It is true that it would be uneconomical to build a machine capable only of ordinary intellectual attainments, but it seems fairly probable that if this could be done then, at double the cost, the machine could exhibit ultraintelligence
- There are ethical problems, such as whether a machine could feel pain especially if it contains chemical artificial neurons, and whether an ultraintelligent machine should be dismantled when it becomes obsolete

- Until an ultraintelligent machine is built perhaps the best intellectual feats will be performed by men and machines in very close, sometimes called "symbiotic," relationship
- In order to achieve the requisite degree of ultraparallel working it might be useful for many of the elements of the machine to contain a very short-range microminiature radio transmitter and receiver
- There is a great waste in having only a small proportion of the components of a machine active at any one time
- Whether a machine of classical or ultraparallel design is to be the first ultraintelligent machine, it will need to be able to handle or to learn to handle ordinary language with great facility. This will be important in order that its instructor should be able to teach it rapidly, and so that later the machine will be able to teach the instructor rapidly
- A man cannot learn more than ten million statements in a lifetime *(source?)*
  - A machine could already store this amount of information without much difficulty, even if it were not ultraparallel, but it seems likely that it would need to be ultraparallel in order to be able to retrieve the information with facility
  - It is in recall rather than in retention that the ordinary human memory reveals its near magic
- For men, meaning serves a function of economy in long-term retention and in information handling
  - This is the basis for our contention that semantics are relevant to the design of an ultraintelligent machine
- In communication, a process of "generalized regeneration" always occur, and that it serves a function of economy

2.3 3. Communication as Regeneration

- The advantage of remembering the word rather than the precise sound is that there is then less to remember and a smaller amount of information handling to do
- This process of regeneration occurs to some extent at each of the levels of phonemes, words, sentences, and longer linguistic stretches, and even at the semantic level, and wherever it occurs it serves a function of economy

2.4 4. Some Representations of "Meaning" and Their Relevance to Intelligent Machines

- When we ask for the meaning of a statement we are talking about language, and are using a metalanguage; and when we ask for the meaning of "meaning" we are using a metametalanguage
- For human beings, meaning is concerned with the outside world or with an imaginary world, so that representations of meaning that are not entirely linguistic in content might be more useful for our purpose
- The behaviorist regards a statement as a stimulus, and interprets its meaning in terms of the class of its effects (responses) in overt behavior
- We call an object a cow if it has enough of the properties of a cow, with perhaps no single property being essential
- An object is said to belong to class C if some function $f(p_1, p_2, ..., p_m)$ is positive, where the p's are the credibilities (logical probabilities) that the object has qualities $Q_1, Q_2, ..., Q_m$

2.5 5. Recall and Information Retrieval

- The usual method for attacking the problem of document retrieval, when there are many documents (say several thousand), is to index each document by means of several index terms
- The process can be made more useful, not allowing for the work in its implementation, if the terms of the documents, and also those of the customer, are given various weights, serving in some degree the function of probabilities. We then have a weighted or statistical system of information retrieval

- When we wish to recall a memory, such as a person's name, we consciously or unconsciously use clues, analogous to index terms
- The speed of neural conduction is much too slow for a primarily serial search to be made

2.6 6. Cell Assemblies and Subassemblies

- A cell assembly is assumed to consist of a great number of neurons, which can all be active at least once within the same interval of about a quarter to half a second
- An assembly reverberates approximately as a unit, and, while reverberating, it tends to inhibit the remainder of the cortex, not neuron by neuron, but enough so that no other assembly can be very active during the same time interval
- It will be assumed that there are also subassemblies that can be active without dominating the whole cortex, and also that when an assembly becomes fatigued and breaks up it leaves several of its own subassemblies active for various lengths of time, from a second to several minutes, and typically about ten seconds
- Each subassembly would consist of a smaller group of neurons than an assembly, but with greater relative interconnectivity
- It might well be that subassemblies correspond to unconscious and especially to preconscious thoughts, in the wakeful state as well as in sleep
- We assume that the strength of a synapse, when not in use, occasionally mutates in the direction of some standard value. This mechanism would explain the gradual erosion of memories that have not been recalled, and would also help to prevent all synapses from reaching equal maximum strength, which would of course be disastrous
- We can distinguish between "reality" and imagination because a memory of a real event is strongly connected to the immediate low-order sensory and motor assemblies. As a memory ages it begins to resemble imagination more and more, and the memories of our childhood are liable to resemble those of a work of fiction
- One of the advantages that an ultraintelligent machine would have over most men, with the possible exception of millionaires, would be that it could record all its experiences in detail, on photographic film or otherwise, together with an accurate time-track. This film would then be available in addition to any brain-like recordings
- A sentence lasting ten seconds would correspond to an assembly sequence of about twenty assemblies
- Which assembly becomes active at the next moment must depend on the current sensory input, the current dominant assembly, and the currently active subassemblies. Indirectly, therefore, it depends on the recent assembly sequence, wherein the most recent assemblies will have the greatest influence
- Primed neurons: After an assembly has just been extinguished, many of its neurons will have received subthreshold activation without having fired. Primed neurons are easy to reactivate during the next few seconds
- The theory of subassemblies is so natural for any large partly random-looking communication network (such as that of a human society) that it tempts one to believe, with Ashby, that a very wide class of machines might exhibit intelligent behavior, provided that they have enough interconnectivity and dynamic states
- That some design is necessary can be seen from one of the objections to the cell assembly theory as originally propounded by Hebb. Hebb did not originally assume that it was necessary to assume inhibition, and Milner pointed out that, without inhibition, the assemblies would fill the whole cortex. Ultimately there could be only one assembly. Either inhibition must be assumed to exist, as well as excitation, or else the assemblies would have to be microscopically small in comparison with the cortex
- There must surely be some advantage in having thin cortices, otherwise people would have thicker ones. It seems unlikely that the brain contains many useless residuals of evolutionary history. Hence the anatomy of the brain is very relevant to the design of the first ultraintelligent machine, but the designer has to guess which features have important operational functions, and which have merely biochemical functions
- The features of a good short-term memory ("attention span"), of the order of 20τ, where τ is the active

time of a single assembly, is certainly essential for intelligence. It might even be possible to improve on the performance of a brain by making the average duration of the sequence somewhat greater than 20τ. But there must be a limit to the useful average duration, for a given cost in equipment
  - It is more likely to be determined by the fact that the complexity of concepts can be roughly measured by the durations of the assembly sequences, and beyond a certain level of complexity the brain would not be large enough to handle the relationships between the concepts
- When guessing what biological features are most relevant to the construction of an ultraintelligent machine, it is necessary to allow for the body as a whole, and not just the brain: an ultraintelligent machine would need also an input (sensorium) and an output (motorium)
- The assembly theory is made easier to accept by combining it with this hypothesis of a central control
  - The greater the amount of activity in the cortex, the greater the number of inhibitory pulses sent to all currently inactive parts of the cortex by the centrencephalic system
    * This negative feedback mechanism would prevent an assembly from firing the whole cortex, and would also tend to make all assemblies of the same order of size, for a given state of wakefulness of the centrencephalic system
- This assembly theory has two merits
  - It would allow a vastly greater class of patterns of activity to assemblies: they would not all have to have the pattern of a three-dimensional fishing net, filling the cortex
  - A single mechanism can explain both the control of the "cerebral atomic reactor" and degrees of wakefulness, and perhaps of psychological "set" also
- It is proposed therefore that our artificial neural net should be umbrella-shaped, with the spikes filling a cone
- During wakefulness, most assemblies will have a very complicated structure, but, during dreamless sleep, the centrencephalic system will become almost exclusively responsible, directly and indirectly, for the activity in the cortex
- Since we are assuming that the duration of a cell assembly is about half a second, following Hebb, it is to be expected that the period of this simple harmonic motion will also be about half a second. This would explain the delta rhythm which occurs during sleep
  - Apparently, very rhythmic assemblies do not correspond to conscious thought
- In order to explain the alpha rhythm, of about five cycles per second, when the eyes are closed and the visual imagination is inactive, along similar lines, we could assume that "visual assemblies" have a duration of only about a fifth of a second
- In the design of an ultraintelligent machine based on an artificial neural net, one of the most vital problems is how to ensure that the mechanism by which an assembly representing a clump of assemblies tend to be formed (the idea of abstraction or generalization) will be effective
  - It seems to be necessary to assume that, when an assembly is active, it causes a little activity in all the assemblies with which it is closely associated, although only one at most of these assemblies will be the next to fire
- It is possible that one of the functions of sleep is to give the brain an opportunity of consolidating the waking experiences by means of unconscious botryological calculations, especially those leading to improved judgments of probabilities
- Human memory levels
  - Immediate recall (about 0.5 second) Concepts currently in consciousness, embodied in the currently active assembly
  - Very short-term memory or attention span (0.5 to 10 seconds) Embodied in the currently active subassemblies, largely the residues of recently active assemblies. The span might be extended up to several minutes, with embodiment in subsubassemblies, etc.
  - Short-term (from about 10 seconds or 10 minutes to about one day) Embodied in primed neurons
  - Medium-term (about one day to about one month, below the age of 30, or about one week above the age of 50) Assemblies are neither partly active nor partly primed, but present only by virtue of their patterns of synaptic strengths, and with little degradation
  - Long-term (about one month to a hundred years) As in medium-term but with more degradation of pattern and loss of detail
- Let us make the following provisional and artificial assumptions

- The probability, in a new brain, that a pair of neurons is connected is the same for every pair of neurons
- Each neuron has μ inhibitory synapses on it, and vastly more excitatory ones
- A single "pulsed" inhibitory synapse dominates any number of pulsed excitatory ones, during a summation interval
- An assembly occupies a proportion α of the cortex and the active subassemblies not in this assembly occupy a proportion $\beta - \alpha$, making a total activity equal to β
- A random neuron has probability $(1 - \beta)^{\mu}$ of escaping inhibition

3 See also

4 References

- Good, Irving John. "Speculations concerning the first ultraintelligent machine." Advances in computers 6 (1966): 31-88.