EVALUATING AND IMPROVING THE IBM COMPUTER PROGRAM 'WATSON'

Michael Scriven
Claremont Graduate University, Western Michigan University
& Palo Alto University

INTRODUCTION A: for the general reader. We'll try here to put some perspective on Watson and its possible progeny, and suggest how to breed the best brood from it.

INTRODUCTION B: **for those particularly interested in program evaluation.** This is a study on the topic of 'not all programs should be judged by the Program Evaluation Standards.' Suppose you are taking an examination to determine your competence in evaluation for a state licensing agency, and you were asked to write a short essay on the above topic. Would you refuse and sue the agency for improper design of the test, on the grounds that computer programs aren't the kind of program that you are trying to prove your competence to evaluate? Or would you try to answer it, and if so, how? The following are some thoughts that occur to me as part of an answer, though they are just first thoughts and can surely be improved by criticism.

CONTEXT A (for the general reader) The general context of the Watson project can be put, roughly speaking, as the search for a winner of the Turing prize, which is offered every year for the best effort at developing a computer program that can pass the Turing test, i.e., be indistinguishable from a human via online interrogation. Big steps along the way have been the efforts to match or exceed human performance in tests of high human intellectual skills in high profile testable tasks—for example, IBM's 'Deep Blue' program, Watson's predecessor, eventually beat the world champion at chess, and Watson has now beaten the world champions at Jeopardy. Does this show Watson is intelligent? No, because being the best rifle shot in the world doesn't mean you're any good with a revolver—intelligence requires a substantial level of *general* problem-solving skill, and these are just two very narrow areas within that vast range. Even though the IBM products beat the very best humans there, they can't yet match humans with an IQ of 90 in many areas outside these narrow foci. And of course Watson is not very cost-effective as a source for intelligent answers across the whole of human knowledge and experience. So, it's a long way from the Turing prize since intelligence is only one of several dimensions on which it would have to match human responses to questions. How can it be improved?

CONTEXT B: for evaluators (the general reader should skip this section) It's common to test people's skills in program evaluation (or some other branch of evaluation), for the purpose of course grading or licensure or job selection, by giving them one or more examples of programs to deal with. This usually means giving them a program description and/or performance data, for a program they have never seen before. Now of course there's a big gap between that and the real world situation in which they will have to practice their evaluation skills, where what they have to be able to do is FIRST to create an *accurate* description and *adequate* basic documentation from a client's request or the name and location of a program, and develop a research design and further performance data from observations, testing, interviews, etc., and only THEN go on to apply the other evaluation and reporting skills. One reason for thinking that graduating with an evaluation major

in a good college masters or doctoral program is a better indicator of evaluation competence than any existing test is that the college programs usually, and certainly should, require some experience working on a complete evaluation from RFP to final report and support, including those first steps.

Now if one is interested in teaching GENERAL evaluation skills, meaning skills that will apply across all sub-areas of evaluation—not just in program evaluation, but also in fields like performance evaluation (which includes educational testing), product evaluation, policy analysis, personnel evaluation, etc., then the appropriate tests must come from sampling from all those domains. Obvious enough, right? Yes, but what's not so obvious is that this kind of testing should include cases where it's not clear which of these domains is home. After all, there are some important differences between program and personnel evaluation, even if the basic logic is the same (see the examples of field-specific checklists on this web site), and you need to decide which tools to use. So, is our example here a case where the checklists (and/or consultants, etc.) from performance evaluation, or product evaluation, or even program or policy analysis are appropriate: or what mix of them? Watson's appearance on the TV show Jeopardy was a performance: are we basically just evaluating that? Or the total abilities of the Watson program? Or the ten million dollar Watson program that IBM set up to achieve this result; or the policy that led to that use of IBM resources?

Since we don't have a client for this evaluation, we can't ask for clarification from that source. Let's just adopt the most practical and socially useful stance: assume we are dong *formative product evaluation*, aiming to help the IBM team improve Watson, in order to provide the company and our society with the most useful tool possible.

TECHNICAL EXPERTISE Isn't this a topic calling for a rather high level of technical expertise in the computer field? It's true, as with most program evaluations, that some area of technical expertise is involved, and you might need to employ a consultant to provide it. But, as usual, there are several dimensions of technical expertise that are relevant here, and what we have with the present author may be enough to support some useful evaluative conclusions, although certainly not all. At the conceptual level—we'll see what that means in a minute—and at one practical level we're in pretty good shape. The present author wrote the first response to the original paper by Alan Turing (the paper in which Turing proposed the Turing test), both being published in the same journal;[1] and some other papers in the early days of artificial intelligence, also reprinted several times.[2] He was called in as one of the small team of consultants the USAF asked to help explain the failure of their

---

[1] Scriven, M. (1953). The mechanical concept of mind. *Mind, 51*, pp. 320–340; reprinted (1963) In Sayre & Crosson, (eds.) *The Modeling of Mind*. University of Notre Dame Press; and (1964). Anderson, A.R. (ed.) *Mind and Machines*. Prentice-Hall

[2] For example: Scriven, M. (1960). The compleat robot: A prolegomenon to androidology. In Hook, S. (ed.) *Dimensions of Mind*. New York University Press, pp. 118–142; reprinted (1970) Feinburg, J. (ed.) *Reason and Responsibility*, 2nd edition; reprinted (1970) In Crowson, F.J. (ed.) *Human and Artificial Intelligence*. Appleton-Century-Crofts, pp. 117–141; reprinted (1970) Kuykendall, E. (ed.) *Philosophy in the Age of Crisis*, Harper & Row, pp. 315–329

mainframe computer (it was called STRETCH and was then the most powerful computer in the world) to provide acceptable accuracy in translating Russian, and wrote the report providing our answer.[3] He was asked to consult for IBM when they faced a critical policy decision about continuing their work on computer-assisted instruction, and gave them advice they followed. On the specific practical task of specifying exactly what tasks a program should be able to perform in order to perform well at a particular task, he authored a book to cover this for the extremely complex task of word processing.[4] And, more recently and at the conceptual level again, he published a rather detailed information-processing (specifically, a neuro-economic) account of the main congitive functions involved in scientific method.[5] Now on the basis of this experience, can we generate any insights about Watson?

THE WATSON SAGA IBM has released some important details of the highlights in the progress from Deep Blue to Watson. There were two non-obvious steps that produced very large improvements in the program's score (measured by running simulations of a Jeopardy session). The obvious steps included dumping vast quantities of trivia and specialized information into Watson's memory, and all the previous Jeopardy questions and answers. The first key step was to abandon the attempt to provide rules to define concepts that humans operate with—their example was the concept of an "A" as represented in a thousand typefaces and graphics—and instead to switch to what is sometimes called 'machine learning' i.e., the mere provision of a huge number of positive and negative examples of the concept, from which the computer extracted its own recognition procedure. (Of course, this is what humans do—what they can't do is to say how they do it—so it might be better called 'natural learning.') The second major payoff step was to provide *in real time* the answers given by the other contestants to questions that Watson got wrong, in the practice trials against experienced and fairly successful Jeopardy graduates.

OPTIMIZING THE FUTURE OF WATSON So now what? IBM already has a number of clients for Watson's services, but how can those and other possible clients be best chosen and best use those services? First, there is a third step that should be taken in dealing with client's questions, from which one can reasonably expect to greatly improve Watson's performance, and which was impossible in the Jeopardy context. This is to allow—in fact, en-

---

[3] Scriven, M. (1964). *Computers and comprehension*. RAND monograph. The answer we gave, albeit not a solution to the problem, was that STRETCH's key flaw was that it operated on purely syntactic rules and could never avoid disastrous errors until it could handle semantic rules as well, i.e., develop a concept of reality as something distinct from descriptions of reality. Solving the problem of how to do that has been a major focus of the revolution in AI/expert systems design that has occurred since then, and is still a long way from completion. The 'Toronto error' by Watson exemplifies exactly the kind of error STRETCH made, one that would be somewhat unfortunate in a computer guiding USAF missiles. (There have been some excellent discussions of this topic in the *NYTimes* since the Jeopardy success.)

[4] Scriven, M. (1987). *Word magic: Evaluating and selecting word processing*. Wadsworth and van Nostrand. Translated into Russian, Moscow: Soviet State Publishing House.

[5] Scriven, M. (May, 1994). The psycho-logical foundations of modern science. *The New Metaphysical Foundations of Modern Science*. Noetic Sciences Institute, pp. 47–79.

courage—Watson to ask questions in order to clarify its understanding of the questions it is being asked. This will cut down—in many situations eliminate—the Toronto kind of mistake. Second, it is important to understand that the Achilles heel in Watson's armor against error, is only a heel. Achilles, heel and all, is still a fearsome warrior, and if aided by a helper who is dedicated to guarding his heels, probably a better warrior than any other. In other words, it doesn't matter that Watson can't pass the Turing test, a team with Watson in it can solve a million problems that no purely human team can handle, and probably earn a thousand dollars a minute of Watson's time. The Turing focus is not the productive focus for practical problems. What kind of problems could it handle? For example, it can certainly handle a huge number of easy life-saving problems like scanning all prescriptions signed by all medical personnel in a chain of hospitals for known drug incompatibilities, and all building specs for large office buildings in Chicago for safety violations.

Those are what we might call Class A problems, and there may be enough of them around to keep Watson busy. But work should go forward on extending its capabilities to Class B problems, graphical recognition problems. For example, with some tweaking of its graphical scanning capabilities, it might be able to do a better job than any human oncological diagnostician in spotting cancer in X-rays, CT scans, and MRIs; but the best approach may be simply to see if it can improve the accuracy of a good specialist to make up a duo that is far better than any human duo. It is not clear that it can handle Class C problems, which are prediction problems e.g., about the future stock market performance of fund holdings. Here the practical problem is not doing very well (which may be impossible because of the noise in the system), but only doing better than, or with, the existing human experts, the fund managers. In general, Watson only has to clear the bar of our best shot in order to be very useful. Then of course there are the Class D problems: e.g., finding a cure for melanoma. Here the issue is as much about segmenting the question as about finding an answer. For it is possible that we already know enough about the chemical and physiological properties of existing substances for a cure to be found amongst them by a better search engine; that Watson could do. But it's at least as likely that the best it could do would be to identify key experiments that need to be done, from whose results it might be possible to infer a possible cure, for further testing. That Watson might be able to do, with ingenious but feasible tweaking. And we should live in hope—enough hope to keep us working on improving Watson, version 69—about Class E problems, e.g., the cure for war. But perhaps we already know how to prevent war. Then the Class E problem becomes how to get us to use the cure. Well, it can't hurt to ask Watson that, too; after all, the bar isn't very high. And we might listen to Watson.