Available online at www.sciencedirect.com

ScienceDirect

Artificial Intelligence 171 (2007) 1183–1186

Artificial Intelligence

www.elsevier.com/locate/artint

Level-headed

Drew McDermott

Yale University, USA

Available online 10 October 2007

Abstract

I don't believe that human-level intelligence is a well defined goal. As the cognitive-science community learns more about thinking and computation, the mileposts will keep changing in ways that we can't predict, as will the esteem we assign to past accomplishments. It would be fun to have a computer that could solve brain teasers as well as the average scientist, but focusing on such things, besides being parochial, overlooks the crucial role language plays in everything humans do, a role we understand hardly at all on a computational level. I am optimistic that we will eventually figure language out, but not without new ideas. Plus, when we can talk to machines, will we understand each other?

Keywords: Speculation; Methodology; Natural language

The question when and how we will attain human-level artificial intelligence is hard to answer because it is so ill-posed. It's like asking a biologist in 1820 how and when Frankenstein's monster would become a reality, replacing Frankenstein with Turing.[1] The history of biology is the history of attacking and solving small technical problems one after another, with many blind alleys. At the end, the goal of knitting parts of dead people together to make a new person just doesn't seem as attractive as it once did.

There's no reason to suppose AI will be any different. Although some old-timers decry our loss of the vision the field exhibited 50 years ago, this sort of loss is exactly what we would hope for. By contrast, phrenology never turned into a real science, and was cranking out the same kind of broadly scoped, superficial, and unverifiable observations at the end as it had at its start.[2]

The phrase "human-level" subtly presupposes that we are measuring skills along one dimension. Humans stand at one altitude, far above fungi and cats, and we are pushing machines up the hill as though moving pianos. But intelligence is the ability to imagine. There are as many different kinds of intelligence as there are kinds of imagination. A computer solving a complex resource-allocation task already surpasses humans in the ability to imagine ways to allocate the resources. Computer programs will eventually have many such skills, but there will never be a time where their total "equals" those of the average human.

I consider it likely that as we solve one technical problem after another—and waste time on several promising ideas that go nowhere—we will eventually get a picture of how computation and thought fit together that we simply can't envisage. When I say "we," I don't mean the AI community by itself, but the entire cognitive-science commu-

E-mail address: drew.mcdermott@yale.edu.

[1] Turing was his own Mary Shelley.

[2] Just spend some time with a few issues of the *American Phrenological Journal*, published from 1838 to 1911.

0004-3702/$ – see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.artint.2007.10.013

nity, defined as all disciplines based on the working hypothesis that the important processes going on in brains are computational. AI gets its ideas from computer science, psychology, linguistics, neuroscience, and occasionally even philosophy. It gives back algorithms and empirical evidence that they work (mathematical proofs of competence being rare in this business).

It seems to me that arguments such as Kurzweil's [1] that exponential growth in computer science is bound to produce superhuman intelligence, and breath-takingly fast as we near that apotheosis, are flawed by the fact that we don't know what we're measuring. Grant that you can fit an exponential curve to scientific output to date in any discipline; will it be the *same* curve from century to century? Whenever there is a scientific revolution, much of what passed for important advances suddenly becomes distracting filler. One now sees the history of that field in terms of an obscure but luminous path leading up to the new insight, which now precedes the so-called "knee" of the exponential that we thought we had already seen.[3] This is not a rare phenomenon. In our own lifetime[4] we have seen a dramatic change in our understanding of learning. In 1950 it seemed that behaviorism was on the verge of explaining the entire human psyche in terms of some simple learning mechanisms. *Time* magazine even adopted the title "Behavior" for its section on human psychology. A graph of the progress of learning theory measured by papers published and number of rats explained would have shown exponential growth. Now we use learning algorithms to sift through huge masses of data, but they work within sharp limits set by computational learning theory. In retrospect, behaviorism didn't even ask the right questions, including most notably: How do the rats compute the relevant properties of the current situation? A revised graph of progress in the field would sharply discount all those rat runners and display quiet exponential growth leading from mathematical logic and empirical linguistics through Chomsky, Valiant, and their successors into a profusion of applications.

When people demand human-level intelligence, they often think in terms of an example such as this one:

In my office I have a device for showing the day of the month. It consists of two little wooden cubes, each of which has a digit inscribed on each face. So there are a total of 12 faces, 6 per cube. To display today's date, you select one cube and face for the tens digit and one for the ones digit. The same cube need not always occupy the same position; sometimes one is in the tens position, sometimes the other. Single-digit days are displayed as $0d$. (The month is shown on a separate set of wooden pieces, which need not concern us.)

I would be very impressed if a computer could prove that such a date-display system was impossible, by the following argument: We're going to need a 0, 1, and 2 on each cube. That's because not all the digits from 1 to 9 will fit on one cube, and 0, 1, and 2 must appear in the tens position opposite each of them. So 0, 1, and 2 must occupy 6 faces. But that leaves just 6 faces for the 7 digits above 2. "QED"

In spite of this proof, I really do have such a date-display system in my office. So there must be a trick. At this point a person might start trying to physically construct the cube, by laying out face arrangements on paper that could be folded into a cube. Actually, a person would almost certainly start trying to construct the cube *before* going off to prove that it was impossible. The lemma that there must be two each of the digits 0, 1, and 2 would be arrived in the process of trying to do the construction, and the constructor would quickly realize that this posed a serious problem.

Presumably, during the construction process the constructor would realize that how the sides folded up was irrelevant. Any assignment of digits to a cube would be as good as any other, so long as there's a 0, 1, and 2 on each cube. On the other hand, they might keep in the mind that they're operating in a weird zone where they "know" the enterprise is doomed, so they might reserve the right to go back and consider the physical arrangement later.

In the process of writing the digits down, the person might somehow realize (don't ask me how) that a 6 and a 9 look very similar, and then realize that we can use one cube face for both, provided we pick a font in which turning the 6 over makes it into a presentable 9. So the seven digits will fit on six faces! After that insight, it will become clear that any assignment of the digits 3–8 to the two cubes will work.

What would it take for a computer to solve this problem? One might argue that the computer would have to be embodied as a robot so that it could write the digits out and see the similarity between 6 and 9. I really don't see why the process wouldn't work just as well in the mind's eye. The important ability the machine must have is to make simplifying assumptions away from the full physical reality of characters stamped on wood, and yet to make good guesses about which of those assumptions to revoke when trouble arises. Two such simplifications are to think of the

[3] Why "knee"? It looks more like an elbow to me. Anyway, it's an optical illusion, of course; draw the curve at a different scale and the knee will appear wherever you want it to.

[4] "Our" here refers to us old-timers.

digits as arbitrary tokens whose only property is to represent a number between 0 and 9, and to assume that one can neglect the orientation of the visible face of a cube. What strikes one as "intelligent" about the ability to solve this problem is the choice of what assumption to question; it wouldn't help to think about the position of the digits on each face, or the possibility of using different colors for different digits, or whether the cubes could be made of cedar or pine.

I did not solve the calendar problem myself; my wife[5] got the little wooden gadget at an office-party gift exchange, and we don't know who created it. Obviously, at least one person figured out how to make it, but who knows what percentage of the human population could? My guess is very few.[6]

People are seldom just presented with problems like this one in cold, precise paragraphs. Instead, they must engage in *conversations* about them, just to get clear on what the problem is. Even if they ask no other questions, after their eureka moment they would surely ask/shout, "Is it okay to use one face to represent both the 6 and the 9?!" I will return to this key point below.

We might suppose that we have not reached human-level intelligence until we have a program that can solve this puzzle, and the "mutilated checkerboard" [2], and some other brain teasers. But why stop there? We could perhaps also demand that it compose a decent sonata, write the chess column for a newspaper, structure the merger of two corporations, *negotiate* the merger of two corporations, play starting quarterback at high-school level, save someone's life by working a suicide hotline, dance the role of Clara in *The Nutcracker*, … .

Obviously, I am deprecating what Seymour Papert long ago called the "superhuman human" fallacy [3], that the success of AI depends on creating programs that do as well as the best humans at various tasks. If we free our minds from this tendency (which we academics are particularly prone to, especially in the brain-teaser department), we are left with the observation that most people use their "human-level" intelligence to *be people*. They use it to solve some fairly well-defined problems that arise in the course of their jobs, their hobbies, foraging for food and entertainment, and worshiping whomever they may or may not worship, but they also use it for managing the interpersonal relationships that are part of all of the above. So to achieve human-level intelligence, we might have to build a person. Perhaps Dr. Frankenstein was right after all.

And back we come to the issue of conversation. The biggest problem the cognitive-science community will face, in my opinion, is language,[7] our Dark Continent. Unless something has happened recently that I am unaware of (which wouldn't be that surprising), we have actually made negative progress in this field in the last twenty-five or so years. That is, we used to have what we thought were reasonable models of "understanding language." We don't anymore. There has been a lot of progress in the theory of syntax, and some interesting ideas about parsing. Speech recognition works well enough for many applications. But if you listen to two or three people talking, and ask, what exactly is each of them trying to do and to what degree are they succeeding?, I don't see what computational answers are currently in contention. Even in a fairly well-structured conversation, such as agreement on the terms of the date-display puzzle, there is a lot going on besides exchanges of formally defined Qs and As, including subtle negotiations about how important it is to Participant Q that Participant A take the problem seriously, how many hints A can negotiate, and so forth. It's not that we have trouble understanding the transactions. Any normal person can join the conversation and pick up on a large fraction of the social signals right away. But just as the ease with which we see things actually makes computer vision harder, the ease with which we interact socially has made us blind, so far, to the possibilities for understanding these interactions in terms of computation.[8]

I am not a pessimist. I think this Dark Continent will be traversed, divided, and conquered, by the same patient, technically guided, interdisciplinary research that has worked thus far. Eventually we will have machines that can carry on conversations. But those machines will have goals and abilities very different from ours, and talking with them will probably not feel like talking to people. The machines will do much better than us at some things and worse at others, and in many areas will not even be competing. By the time we're in a position to know whether they've reached our level, the concept of "level" will surely be obsolete.

[5] Whom I also thank for editing this paper.

[6] Another guess is that most scientists would eventually arrive at the solution; this conjecture is corrupted by vanity, though.

[7] I mean "natural" language, of course, but the qualifier hardly seems necessary in the context of this discussion.

[8] Reading an essay, or listening to a recorded essay, with no chance for the reader to talk back, seems like a simpler problem. Maybe. But when I read I find myself continually wanting to talk back to the writer, and having to imagine how he or she would reply. Is this fantasy conversation easier to deal with than the real thing?

The foregoing paragraph is, of course, pure speculation. It's entirely possible that cognitive science will someday seem as wrong-headed as phrenology. Plenty of skeptics find our intuition ridiculous that computation is powerful enough to explain almost everything about intelligence, and the burden of proof is still on us. However, for me and many others it is too strong an intuition to ignore.

Assuming that computationalism is eventually borne out, the consequences are impossible to predict, as the consequences of future developments always have been. So I'll make an observation about the present. We already have computers making decisions for us. It would be out of the question to hire a bunch of people to manage the traffic in electronic networks, to schedule airline flights, to manage the logistics of military campaigns, to keep track of the interactions among a hospital's patients' medications.[9] Our lives are affected almost continuously by computerized decisions. By the time the computers can chat with us, it's reasonable to assume that this process will have expanded even further, with almost all our decisions monitored, influenced, or just made for us by machines. Will the intelligent ones be able to explain what's going on and justify themselves to us? Or will they apologize and say that they can't introspect about the computational processes squeezing and tugging at our lives any more deeply than we can introspect about digestion? Will it make any difference?

References

[1] Ray Kurzweil, The Singularity is Near, Viking, New York, 2005.

[2] John McCarthy, A tough nut for proof procedures, Technical Report 16, Stanford AI Project, 1964.

[3] Pamela McCorduck, Machines Who Think, Freeman, San Francisco, 1979.

[9] And we find it morally repugnant to persuade a person to carry a bomb into an area crowded with civilians when a silicon chip can do it just as well.