The Singularity and the State of the Art in Artificial Intelligence

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

October 13, 2013

Abstract

The state of the art in automating basic cognitive tasks, including vision and natural language understanding, is far below human abilities. Real-world reasoning, which is an unavoidable part of many advanced forms of computer vision and natural language understanding, is particularly difficult. This suggests that the advent of computers with superhuman general intelligence is not imminent. The possibility of attaining a singularity by computers that lack these abilities is discussed briefly.

When I was invited to contribute an article on the subject of "The Singularity", my initial reaction was "I fervently pray that I don't live to see it." Parenthetically, I may say the same of the immortality singularity. I am 55 years old, as of the time of writing, and, compared to most of humanity, I live in circumstances much more comfortable than anything I have earned; but I am hoping not to be here 50 years hence; and certainly hoping not to be here 100 years hence. Whether I will view this question with the same *sang froid* as the time comes closer remains to be seen.

However, my personal preferences are largely irrelevant. What I primarily want to do in this essay on the singularity is to discuss the one aspect of the question where I can pretend to any kind of expertise: the state of the art in artificial intelligence (AI) and the challenges in achieving human-level performance on AI tasks. I will argue that these do not suggest that computers will attain an intelligence singularity any time in the near future. I will then, much more conjecturally, discuss whether or not an intelligence singularity might be able to sidestep these challenges.

1 Artificial Intelligence

What is artificial intelligence? There are numerous different definitions, with different slants (see [9] chap. 1 for a review and discussion). However, the definition that I prefer is this: There are a number of cognitive tasks that people do easily — often, indeed, with no conscious thought at all — but that it is extremely hard to program on computers. Archetypal examples are vision, natural language understanding, and "real-world reasoning"; I will elaborate on this last below. Artificial intelligence, as I define it, is the study of getting computers to carry out these tasks.

Now, it can be objected that this definition is inherently unfair to the computers. If you define AI as "problems that are hard for computers," then it is no surprise that these are hard for computers.

We will return to this point below, in section 2. For the time being, though, I think it can be agreed, that, whether or not these abilities are *necessary* for a super-intelligent computer, they would certainly be an *asset*; and therefore, in speculating on the singularity, it is at least somewhat relevant to consider how well these tasks can now be done, and what are the prospects for progress.

Some forms of computer vision and natural language processing can currently be done quite well. In vision: good-quality printed text can be converted to electronic text with error rates that are very small, though low-quality or old documents can be more problematic. The current generation of ATM's can read handwritten checks. As long ago as 1998, an autonomous vehicle drove across country[1]; the current generation of autonomous vehicles do quite well on the much more challenging problem of off-road driving. For certain kinds of technical image analysis, (e.g. in medical applications), computers do as well or better than human experts. In natural language: automated dictation systems work with an error rate of 1-5% under favorable circumstances. Web search engines find relevant Web documents with impressive accuracy and astonishing speed. Google Translate produces translations that are sometimes excellent and generally good enough to be useful, though usually with some rather strange-looking errors.

However, it is critical to understand how far the state of the art is from human-level abilities. The point is best illustrated by example. The following are representative of cutting-edge research; they are taken from papers at the top research conferences[2] in computer vision and natural language processing in 2010-2011.

Recognizing birds in images. A program to recognize major components of a bird's body (body, head, wings, legs) and to identify the category of bird (duck, heron, hawk, owl, songbird) in pictures of birds achieved a precision of about 50% on finding the components, and of about 40% on identifying the category [5].

Identifying images that match simple phrases.. A program was developed to identify images in a standard collection that match simple phrases. This was very successful for some phrases; e.g. at a 50% recall cutoff, the precision was about 85% for "person riding bicycle" and 100% for "horse and rider jumping". For other phrases, it was much less successful; e.g. at 50% recall, the precision was below 5% for "person drinking [from] bottle", "person lying on sofa", and "dog lying on sofa" [10].

Coreference resolution. A state of the art system for coreference resolution — identifying whether two phrases in a text refer to the same thing or different things — achieved success rates ranging from 82% recall and 90% precision to 40% recall and 50% precision, depending on the source of the text and the grammatical category involved [6].

Event extraction. A program for identifying events of a specified type in news articles; specifically, for identifying the event trigger, the arguments, and their roles. For example, in the sentence "Bob Cole was killed in France today", the trigger for the event `Die` is "killed", the arguments are "Bob Cole" and "France" and the roles are `Victim` and `Place` respectively. There are 33 different event types. The system achieved an F-score (harmonic mean of recall and precision) of 56.9% on trigger labelling, 43.8% on argument labelling and 39.0% on role labelling [8].

Thus, on these simple, narrowly defined, AI tasks, which people do easily with essentially 100% accuracy, current technology often does not come anywhere close. My point is not in the least to denigrate this work. The papers are major accomplishments, of which their authors are justly proud; the researchers are top-notch, hard-working scientists, building on decades of research. My point is

[1] From a technical standpoint, the vision problem here is surprisingly easy; pretty much all the machine has to see are the boundaries of the road, the lane markers, and the other vehicles.

[2] In these areas of computer science, conference publication is as prestigious as journal publication, and prompter. Also, whereas there are areas of computer science research in which it is known or widely believed that much of the most advanced research is being done in secret in government or industrial labs, this does not seem to be the case for vision or natural language processing.

that these problems are very very hard.

Moreover, the success rates for such AI tasks generally reach a plateau, often well below 100%, beyond which progress is extremely slow and difficult. Once such a plateau has been reached, an improvement of accuracy of 3% — e.g. from 60% to 62% accuracy — is noteworthy and requires months of labor, applying a half-dozen new machine learning techniques to some vast new data set, and using immense amounts of computational resources. An improvement of 5% is remarkable, and an improvement of 10% is spectacular.

It should be also emphasized that the tasks in these exercises each has a quite narrowly defined scope, in terms of the kinds of information that the system can extract from the text or image. At the current state of the art, it would not be reasonable even to attempt open-ended cognitive tasks such as watching a movie, or reading a short story, and answering questions about what was going on. It would not even be plausible as a project, and there would be no meaningful way to measure the degree of success.

1.1 Real world reasoning

One of the hardest categories of tasks in artificial intelligence is automating "real-world reasoning". This is easiest to explain in terms of examples. I will start with examples of scientific reasoning, because the importance of scientific reasoning for a hyper-intelligent computer is hard to dispute.

The Wolfram Alpha system[3] is an extraordinary accomplishment, which combines a vast data collection, a huge collection of computational techniques, and a sophisticated front-end for accepting natural language queries. If you ask Wolfram Alpha "How far was Jupiter from Mars on July 16, 1604?" you get an answer within seconds — I presume the correct answer. However, it is stumped by much simpler astronomical questions, such as

1. When is the next sunrise over crater Aristarchus?
2. To an astronomer near Polaris, which is brighter, the sun or Sirius?
3. Is there ever a lunar eclipse one day and a solar eclipse the next?

All that Wolfram Alpha can do is to give you the facts it knows about crater Aristarchus, Polaris, Sirius, and so on, in the hopes that some of it will be useful. Of course, the information and the formulas needed to answer these questions are all in Wolfram Alpha's database, but for these questions, it has no way to put the data and the formulas together; it does not even understand the questions. (The program echoes its understanding of the question — a very good feature.) The fundamental reason is that, behind the data and formulas, there is no actual understanding of what is going on; Wolfram Alpha has no idea that the sun rises on the moon, or that the sun can be viewed from astronomically defined locations. It does not know what a sun rise is or what the moon is; it just has a data collection and a collection of formulas.

Now of course it would be easy — I presume a few hours of work — for the Wolfram Alpha people to add categories of questions similar to (1) or (2) if they thought there was enough demand for these to make it worthwhile; the program could easily be extended to calculate the rising and setting of any heavenly body from any geographic location on the moon, plus its angle in the sky at any time; or to tell the apparent brightness of any star as seen from any other star. However, there would still be endless simple questions of forms that they had not yet thought of or bothered with. Question (3), the easiest for the human reader, is much more difficult to automate because of the quantification over time; the question does not have to do with what is true at a specified time, but with what is

[3] http://www.wolframalpha.com

true at all times. In the current state of the art it would certainly be difficult, probably impossible for practical purposes, to extend Wolfram Alpha to handle a reasonable category of questions similar to question (3).

Once one gets outside the range of scientific reasoning and into the range of everyday life, the problem of reasoning becomes much easier for people and much harder for computers. Consider the following three questions:

4. There is a milk bottle in my field of view, and I can see through it. Is it full or empty?
5. I poured the milk from the bottle to a pitcher, and the pitcher overflowed. Which is bigger, the bottle or the pitcher?
6. I may have accidentally poured bleach into my glass of milk. The milk looks OK, but smells a little funny. Can I drink the milk, if I'm careful not to swallow any of the bleach?

There are no programs that can answer these, or a myriad similar problems. This is the area of *automated commonsense reasoning;* despite more than fifty years of research, only very limited progress has been made on this.[3]

1.2 Real-world reasoning as an obstacle to other AI tasks

Real-world reasoning is not only important as an end in itself, but it is an important component of other AI tasks, including natural language processing and vision. The difficulties in automating real-world reasoning therefore sets bounds to the quality with which these tasks can be carried out.

The importance of real-world knowledge for natural language processing, and in particular for disambiguation of all kinds, is very well-known; it was discussed as early as 1960 by Bar-Hillel [1]. The point is vividly illustrated by Winograd schemas [7]. A Winograd schema is a pair of sentences that differ in one or two words, containing an referential ambiguity that is resolved in opposite ways in the two sentences; for example, "I poured milk from the bottle to the cup until it was [full/empty]". To disambiguate the reference of "it" correctly — that is, to realize that "it" must refer to the cup if the last word of the sentence is "full" and must refer to the bottle if the last word is "empty" — requires having the same kind of information about "pouring" as in question (5) above; there are no other linguistic clues. Many of the ambiguities in natural language text can be resolved using simple rules that are comparatively easy to acquire, but a substantial fraction can only be resolved using a rich understanding of the subject matter in this way.

Almost without exception, therefore, the language tasks where practically successful programs can be developed are those that can be carried out purely in terms of manipulating individual words or short phrases, without attempting any deeper understanding. Web search engines, for example, essentially match the words of the query against the words in the document; they have sophisticated matching criteria and sophisticated non-linguistic rules for evaluating the quality of a web page. Watson, the automated Jeopardy champion, in a similar way finds sentences in its knowledge base that fit the form of the question. The really remarkable insight in Watson is that Jeopardy can be solved using these techniques; but the techniques developed in Watson do not extend to text understanding in a broader sense.

The importance of real-world knowledge for vision is somewhat less appreciated, because in interpreting simple images that show one object in the center, real-world knowledge is only occasionally needed or useful. However, it often becomes important in interpreting complex images, and is often unavoidable in interpreting video. Consider, for example, the photograph[4] in figure 1 of Julia Child's

[4] Photograph by Matthew Bisanz; reproduced under a Creative Commons licence. http://en.wikipedia.org/wiki/File:Julia_Child's_kitchen_by_Matthew_Bisanz.JPG

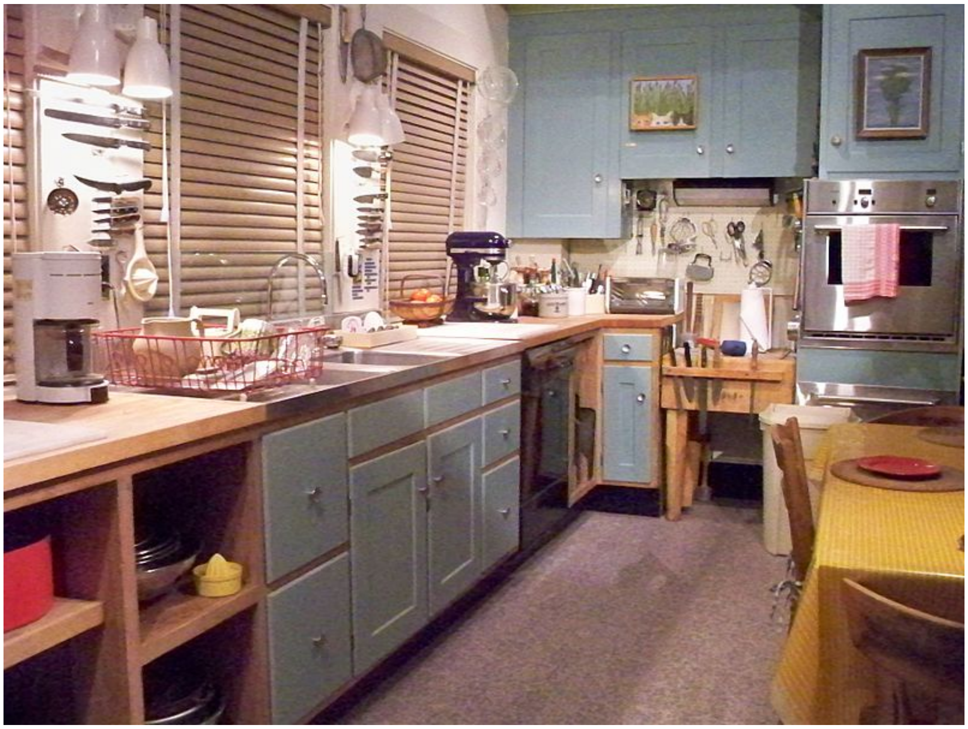

Figure 1: Julia Child's kitchen

kitchen, now enshrined at the Smithsonian Institute. Many of the objects that are small or partially seen, such as the metal bowls in the shelf on the left, the cold water knob for the faucet, the round metal knobs on the cabinets, the dishwasher, and the chairs at the table seen from the side, are only recognizable in context; the isolated image would be hard to identify. The metal sink in the counter looks like a flat metal plate; it is identified as a sink, partly because of one's expectations of kitchen counters, partly because of the presence of the faucet. The top of the chair on the far side of the table is only identifiable because it matches the partial view of the chair on the near side of the table.

The viewer infers the existence of objects that are not in the image at all. There is a table under the yellow tablecloth. The scissors and so on hanging on the board in the back are presumably supported by pegs or hooks. There is presumably also a hot water knob for the faucet occluded by the dish rack.

The viewer also infers how the objects can be used (sometimes called their "affordances"). She knows that the cabinets and shelves can be opened by pulling on the handles; and she can tell the difference between the shelves, which pull directly outward, and have the handle in the center, and the cabinets which rotate on an unseen hinge, and have the handle on the side.

The need for world knowledge in video is even stronger. Think about some short scene from a movie with strong visual impact, little dialogue, and an unusual or complex situation — the scene with the horse's head in *The Godfather*, say, or the mirror scene in *Duck Soup*, or the scene with the night-vision goggles in *The Silence of the Lambs* — and think about the cognitive process that are involved if you try to explain what is happening. Understanding any of these is only possible using a rich body of background knowledge.

1.3 Summary of the present and its implications for the future

To summarize the above discussion:

- For most tasks in automated vision and natural language processing, even quite narrowly defined tasks, the quality of the best software tends to plateau out at a level considerably below human abilities, though there are important exceptions. Once such a plateau has been reached, getting further improvements to quality is generally extremely difficult and extremely slow.
- For more open-ended or more broadly defined tasks in vision and natural language, no program can achieve success remotely comparable to human abilities, unless the task can be carried out purely on the basis of surface characteristics.
- The state of the art for automating real-world reasoning is extremely limited, and the fraction of real-world reasoning that has been automated is tiny, though it is hard to measure meaningfully.
- The use of real-world reasoning is unavoidable in virtually all natural language tasks of any sophistication and in many vision tasks. In the current state of the art, success in such tasks can only be achieved to the extent that the issues of real-world reasoning can be avoided.

Let me emphasize that the above are not the bitter maunderings of a nay-saying pessimist; this is simply the acknowledged state of art, and anyone doing research in the field takes these as a given.

What about the future? Certainly, the present does not give very good information about the future, but it is all the information we have. It is certainly possible that some conceptual breakthroughs will entirely transform the state of the art, and lead to breathtaking advances. I do not see any way to guess at the likelihood of that. Absent that, it seems to me very unlikely that any combination of computing power, “Big Data”, and incremental advances in the techniques we currently have will give rise to a radical change. However, that certainly is a debatable opinion and there are those who disagree. One can muster arguments on either side, but the unknowns here are so great that I do not see that the debate would be in any way useful. In short, there is little justification for the belief that these limitations will not be overcome in the next couple of decades; but it seems to me that there is even less justification for the belief that they will be.

My own view is that the attempt to achieve human-level abilities in AI tasks in general, and to automate a large part of real-world reasoning in particular, must at this point be viewed as a high-risk, high-payoff undertaking, comparable to SETI or fusion reactors.

2 Side-stepping AI

There is also the possibility that some kind of singularity may take place *without* having computers come close to human-level abilities in such tasks as vision and natural language. After all, those tasks were chosen specifically because humans are particularly good at them. If bees decided to evaluate human intelligence in terms of our ability to find pollen and communicate its location, then no doubt they would find us unimpressive. From that point of view, natural language, in particular, may be suspect; the fact that one self-important species of large primate invests an inordinate fraction of its energies in complicated chattering does not mean that all intelligent creatures should necessarily do likewise.[5]

[5] The best discussion of this that I have seen is Clarence Day’s *This Simian World* [4].

To some extent this depends on what kind of singularity is being discussed. One can certainly imagine a collection of super-intelligent computers thinking and talking to one another about abstract concepts far beyond our ken without either vision, natural language, or real-world reasoning.

However, it seems safe to say that most visions of the singularity involve some large degree of technological and scientific mastery. Natural language ability may indeed be irrelevant here, useful only for communicating to such humans as remain. The ability to interpret a visual image or video seems clearly useful, though an ability substantially inferior to people's may suffice. However, an understanding of the real world somewhat along lines of people's understanding would certainly seems to be a *sine qua non.* As far as we know, if you do not have the conceptual apparatus to be able to answer questions like, "Is there ever a lunar eclipse on one day and a solar eclipse on the next?" then you certainly cannot understand science, and almost certainly you will be limited in your ability to design technology. Now it is *conceivable*, I suppose, that post-singularity computers will have some alternative way of approaching science that does not include what we would consider "real-world understanding" but nonetheless suffices to allow them to build technological wonders; but it does not seem likely.

Other possible avenues to superintelligence have been suggested.[6] A machine of powerful general intelligence could itself figure out how to carry out AI tasks. Or a machine of lesser general intelligence could first figure out how to make itself smarter, and then when it got smart enough, figure out how to do AI tasks. All I can say about these scenarios is that I have not see anything in the machine learning literature that suggests that we are anywhere close to that, or headed in that direction. Or with advances in neuroscience, it might become possible to build an emulation of the human brain; and then if we make it a couple of orders of magnitude faster or larger, we have a superbrain. Apparently, we may be fairly close to being able to do this in terms of simple computing power; the main gap is in the neuroscience. Again, this is possible, but it does not seem to be at all imminent.

Yet another possibility would be to have a "human in the loop", along the lines of amazon.com's "Mechanical Turk".[7] One can imagine a civilization of computers and people, where the computers think of people as a kind of seer; in most things, quite stupid and laughably slow, but possessed of a strange super-computer ability to interpret images, and of a mysterious insight which they call "real-world reasoning". Actually, I'm not sure I can imagine it, but I can imagine that someone could imagine it. That second-order fantasy seems to me as good a stopping point as any for this essay.

References

[1] Y. Bar-Hillel, The Present Status of Automatic Translation of Languages, In F.L. Alt (ed.) *Advances in Computers,* Vol. I, Academic Press, 1960, 91-163.

[2] J. Bardin, From Bench to Bunker: How a 1960s discovery in neuroscience sparked a military project, *The Chronicle of Higher Education,* July 9, 2012. http://m.chronicle.com/article/article-content/132743/

[3] E. Davis and L. Morgenstern, Progress in Formal Commonsense Reasoning, *Artificial Intelligence,* **153**(1-2), 2004, 1-12.

[4] C. Day, *This Simian World,* Knopf, 1920.

[6] My thanks to Aaron Weiner for pointing these out.

[7] An extraordinary instance of this, perhaps a harbinger of things to come, is a project reported in [2]. A subject in a lab is shown satellite pictures of desert landscape at the rate of 20 per second. When one of these images shows a building, there is a spike in brain activity that is detected by an EEG.

[5] R. Farrell et al., Birdlets: Subordinate Categorization using Volumetric Primitives and Pose-Normalized Appearance, *13th International Conference on Computer Vision*, 2011.

[6] F. Kong et al., Dependency-driven Anaphoricity Determination for Coreference Resolution, *23rd International Conference on Computational Linguistics (COLING)*, 2010.

[7] H. Levesque, E. Davis, and L. Morgenstern, The Winograd Schema Challenge, *Principles of Knowledge Representation and Reasoning (KR),* 2012.

[8] S. Liao and R. Grishman, Can Document Selection Help Semisupervised Learning? A Case Study On Event Extraction, *Proc. Association for Computational Linguistics,* 2011.

[9] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach. 3rd edn.* Prentice Hall, 2009.

[10] M. Sadeghi and A. Farhadi, Recognition using Visual Phrases *Computer Vision and Pattern Recognition*, 2011.