Wise machines?

Colin Allen and Wendell Wallach

Colin Allen is based in the Department of History & Philosophy of Science and Program in Cognitive Science, Indiana University, Bloomington, Indiana, USA. Wendell Wallach is based in the Interdisciplinary Center for Bioethics, Yale University, New Haven, Connecticut, USA.

Abstract

Purpose – In spite of highly publicized competitions where computers have prevailed over humans, the intelligence of computer systems still remains quite limited in comparison to that of humans. Present day computers provide plenty of information but lack wisdom. The purpose of this paper is to investigate whether reliance on computers with limited intelligence might undermine the quality of the education students receive.
Design/methodology/approach – Using a conceptual approach, the authors take the performance of IBM's Watson computer against human quiz competitors as a starting point to explore how society, and especially education, might change in the future when everyone has access to desktop technology to access information. They explore the issue of placing excessive trust in such machines without the capacity to evaluate the quality and reliability of the information provided.
Findings – The authors find that the day when computing machines surpass human intelligence is much further in the future than predicted by some forecasters. Addressing the problem of dependency on information technology, they envisage a technical solution - wiser machines which not only return the search results, but also help make them comprehensible - but find that although it is relatively simple to engineer knowledge distribution and access, it is more difficult to engineer wisdom.
Practical implications – Creating computers that are wise will be difficult, but educating students to be wise in the age of computers may also be quite difficult. For the future, one might explore the development of computer tools that demonstrate sensitivity to alternative answers to difficult questions, different courses of action, and their own limitations. For the present, one will need to train students to appreciate the limitations inherent in the technologies on which they have become dependent. Originality/value – Critical thinking, innovation, and wisdom require skills beyond the kinds of answers computers give now or are likely to provide in the coming decade.
Keywords Wisdom, Artificial intelligence, Education, IBM's Watson, Frame problem, Computers
Paper type Conceptual paper

When IBM's Watson triumphed over two top Jeopardy competitors in February 2011, the media buzzed with talk of artificial intelligence just as they had 14 years earlier when Watson's predecessor, IBM's Deep Blue, won its rematch with world chess champion Gary Kasparov. Journalists and radio hosts were asking when will computing machines surpass human intelligence? Or had it already happened?

Perhaps some eager college administrator immediately considered the savings from installing Watson on campus. If it could answer student questions as well as it deals with Jeopardy probes, just think of all the classroom space that could be freed by sending students online instead of to discussion sections! Watson is an impressive engineering feat, but it is not yet time to assume that instructors can step away from the lectern.

Will that time ever arrive? Perhaps one day, but we believe that day is much further in the future than is imagined by forecasters such as Vernor Vinge and Ray Kurzweil. They have promoted the idea of an imminent technological singularity, when machine intelligence will exceed that of humans and accelerate beyond our ability to predict and understand it (Vinge, 1993; Kurzweil, 2005). If the singularity comes, we suspect it will be well beyond our lifetimes. A bigger immediate concern of ours is how society and especially its incubator, education, might change in the future where everyone has access to Watson-like oracles on the desktop. Or, as seems more likely, accessible from the cloud by every kind of networked device.

There is already a passionate debate about whether open access to online encyclopedias – promising instant access to the collective knowledge of humanity – enhances or detracts from education. Some faculty and institutions have been moved to ban students from using Wikipedia altogether, while simultaneously supporting more traditional scholarly endeavors that seek to make all manner of research available online. Almost anyone teaching on a college campus has stories to tell about the cut-and-paste attitude towards ‘‘research’’ that seems ever more prevalent among undergraduates. Watson-like natural-language interfaces to the ever growing databases of academic material will only make it easier for students to find what they need when they need it, and to merge it into whatever report they are writing without bothering to digest it fully. This is ‘‘ScholarShip 2.0’’.

Do not get us wrong: We find the technology wondrous and it can surely facilitate collective human achievement. But like any developing organism, it will need to be nurtured carefully if its full potential is to be released. Machines that seem to know a lot are already trusted too much by people who do not know how to evaluate what they find. That is the essence of the Wikipedia problem in education. One might envisage a technological solution: wiser machines which do not just return the search results, but which also help make them comprehensible. But here is the rub. To engineer knowledge distribution and access is relatively simple. To engineer wisdom is not.

What do we mean by this? Think of it this way: Philosophy, a word whose very root means ‘‘love of wisdom’’, has as its first pivotal figure a crotchety fellow who made it his business to show that the others around him did not really know as much as they thought they knew, and that they were not wise at all. This was a thankless task for which he was eventually tried and executed. Wisdom, according to Socrates, comes from an appreciation of how much you do not know. But given that Socrates’ compatriots were not particularly open to the message, how much harder would it be to get a machine to appreciate what it does not know?

The primary new computer skill displayed by Watson lay in deciphering what the question was in order to pursue a relatively straightforward search for factual information. But the publicity following Watson’s Jeopardy triumph also made much of Watson’s internal confidence indicator. Because Jeopardy penalizes players heavily for wrong answers, an over-confident player can quickly end up with a negative score. (So too can a panicking player, but emotions are not part of Watson’s program – another advantage to the machine according to some, although we are not so sure.) Watson’s algorithms for pressing the buzzer included an assessment of the relative likelihood of its three best candidate answers; only if the best of the three exceeded a specific threshold did Watson attempt to buzz in. This was evidently quite effective. Stephanie Kovalchik, a graduate student in the biostatistics department at UCLA, analyzed Watson’s responses for an article titled ‘‘How wise was Watson?’’ (Kovalchik, 2011). She found that Watson was incorrect only 10 per cent of the time it attempted to buzz in. On the 26 occasions when it was below threshold, its best guess would have been correct almost 20 per cent of the time. Kovalchik concludes that Watson could have afforded to be a bit more confident.

Irrespective of that, Watson's actual settings allowed it to do well enough in the game. But game shows are closed worlds with trivial consequences. Furthermore, each mistaken response, whether a false positive (a wrong answer delivered) or false negative (a correct answer undelivered) is effectively independent. The same question will not come up again, and the questions are usually arranged so that there is nothing to be learned from one answer about the next. So, the machine does not have to learn anything from its failures, although it could perhaps adjust its response threshold to optimize its ratio of false positives to false negatives.

There is no indication that Watson made any such adjustments during the game to its algorithms, but even if it did, that would fall a long way short of the kind of self knowledge, metacognitive control, and reflection-driven learning that humans display in the real world.

These capacities involve more than an internal gauge of confidence. They depend on an active search for information outside one's own memory banks and a process of creatively making sense of the information that is gathered. External information gathering is not allowed in Jeopardy. In a limited fashion, it is allowed in Who Wants to be a Millionaire? and contestants even may demonstrate a bit of wisdom in choosing whether to consult a friend or the audience for the particular question at hand. But anything approaching the talents required in high-level research is absent from these contests – it would make for rather slow television, anyway!

In real-world situations, there is no game show host who knows the answer. The world is open, and the solution to a given problem may not be known by anyone. Indeed, researchers are often confronted with incomplete, inaccurate, and contradictory information. People must often act under risk and uncertainty (Gigerenzer, 2008). In the end, students and professionals must decide when to curtail their search and wrap up the report. Wisdom lies not just in knowing one's limits, not just in knowing what one does not know, but also in knowing when and where a bit more effort might pay off. In other words, humans make value judgments that go beyond assessing the accuracy of a search for a factual piece of information when they research the answer to an open-ended question.

In our book Moral Machines (Wallach and Allen, 2009) we assessed the challenge of building artificial moral agents (AMAs). We started with the observation that hardware robots that roam in human-occupied spaces and software "bots" that roam in virtual environments are being designed and released to operate in ever more open environments, with less and less real-time oversight by human beings. In this respect, machines are becoming more autonomous. But they are, as we say in the book, ethically blind. The computers that approve or deny millions of credit card transactions every minute are autonomous and ethically blind. They have no information about the possible effects of the decision to approve or deny on the welfare of the person requesting the charge. Nor do they know that they do not know these things. They have no means to find out what they do not know. And, most importantly, they have no conception of why it might matter to do so.

We did not tackle the topic of wisdom in our book, but we noted that ethics has both a reactive and a deliberative part. The well-schooled individual has not just the tendency to do the right thing, but when that tendency fails, as it inevitably will, she has the means to reflect on the failure, learn from it, and with a bit of luck, do better next time. Whether talking about moral agents or epistemic agents, the virtues, as other philosophers have noted, are rather similar. A lack of complacency, a capacity for independent thought, and a recognition of one's limits as an individual all feed into wise choices, whether they concern ethics or justified belief.

‘‘Ignorance is bliss’’, it is said. But bliss has a habit of ending abruptly and painfully. Intelligent agents readjust. People have the means to figure out what went wrong; computers, not so much. The pain is perhaps an important motivator; machines do not yet have such feelings or emotions. When Watson gets an answer wrong its programmers scramble to make sure it does not happen next time. Watson itself just sits there like the big lump of plastic and metal that it is. People notice patterns, they are aware of their potential significance, and then exploit them. Watson is oblivious to the wider context in which it operates (although Watson’s descendants are not necessarily doomed to be). In questioning whether Watson tells us anything much about human cognition, Doug Hofstadter (pers. comm.) suggests that it would not notice anything unusual if it got the following sequence of prompts:

1. The principal sidekick of a nonexistent person whose fanatical followers the world around proudly call themselves ‘‘The Baker Street Irregulars’’.

2. The first head of a US firm that used Hollerith cards to help businesses.

3. The junior member of the pair of people who first understood the coiling shape of the invisible substance that carries hereditary information in living organisms.

4. A device that was designed to compete against very accomplished human performers in rapid-fire answering of questions about what are commonly called ‘‘trivia’’.

The point is not that Watson could not be programmed to notice repetitions among the correct responses, but that it lacks the kind of enquiring, responsive, flexible intelligence to notice this pattern spontaneously. Adding a literal repetition detector would not do anything to remove the fact that Watson is, in fact, a giant bore with no interest in any of the facts it has been programmed to spit out. This is one central lack in computing machines that still no one knows how to engineer in.

And this is why Watson will not be replacing good teachers or wise counselors any time soon. People who contribute to moral and social development, who shape the society for years to come, know when to joke, when to cajole, when to be compassionate, when to be firm, when to digress, and, above all, when to say ‘‘I don’t know’’ – but without just leaving it at that. They go on to demonstrate how to search for an answer and evaluate it. That means not just stopping with the first answer you find online (whether it be in Wikipedia or the Stanford Encyclopedia of Philosophy). It means being a critical consumer of information, not a clever regurgitator of facts that have been stuffed in from the outside. Good teachers and wise counselors do not just feed their students answers to the hardest questions, but they encourage an inquiring attitude. These are characteristics that our society fails to impart to as many as it should. In education, we may recognize good, wise teaching when we see it, but we do not really have a social technology for reliably producing good, wise teachers. And given that we barely understand what we are doing ourselves, it is not surprising that engineers or philosophers do not yet know how to give these qualities to machines. Maybe it is possible, maybe it is not. We firmly believe there are no sound a priori arguments to say that it cannot be done, or that it can. But whether or not the attempt to build ever more human-like machines ultimately succeeds, we also believe that a lot will be learned in the effort, and the effort will also force human beings to learn a lot more about themselves.

There are some who would call for a moratorium on the kind of research we embrace because they fear a machine uprising, Terminator-style. Others would end it because they fear, Frankenstein-style, for the psyches of the machines themselves. Both of these fears are futuristic in the extreme. There is a third class of fears that have to do with the de-humanizing effects that the machines will have on us. Here the worries have more purchase. How many times have you heard, ''I'd like to help you but the computer won't let me''? More subtly, when everything from quark physics to the latest blockbuster can be made to appear magically on your desktop, it conveys nothing of the individual and collective human effort that was required to generate that iota of knowledge. Of course the toil behind the item can be appreciated and information that is easily accessed can be used wisely to shape further endeavors, enabling new discoveries. But with so many answers to be found in just a click or two, it is easy to worry that the effect on human effort towards exploring harder questions is a net negative.

One concern is whether people will treat the answers so readily provided by computers as authoritative or approach those answers critically. Of course, the old adage ''Don't believe everything you read in the newspaper'' has its digital counterpart, but the seemingly impersonal manner in which computers deliver information may falsely suggest objectivity. Another concern is about what happens when answers to difficult questions do not come quickly. Might some students too quickly just drop the enquiry and go on to a different question? Or (as recently experienced by one of us) when the customer's problem cannot be solved on this screen or the next, might the person in the call center simply invent an explanation in order to get on to the next customer? It is an empirical question whether these kinds of outcomes will become more frequent when people are immersed from an early age in an environment where computers stand ready to answer almost any question 24x7. It is an empirical question that cannot, however, be easily answered experimentally. In a loose sense, the experiment has begun; but it is not being run with adequate controls. Nor could it be.

Creating computers that are wise will be difficult, but educating students to be wise in the age of computer may also be quite difficult. The discussion, then, seems to be thrown back on tales about students who are worse prepared than ever for their educations. Were professors not saying the same thing 50, 100 years ago? These laments about how things have changed for the worse are perhaps more nostalgia than accurate recollection. Things are certainly different. In the developed world, almost all of today's students have grown up with the Internet at their fingertips. Their intelligence is measured, in large part, by their ability to get what they need from their environment, and that environment is ubiquitous network computing. There will be Nobel Prize winners who have never stepped into a traditional library, and with good reason given that for many purposes the online library is already a superior resource. And anyway, let's face it, most of the students in the days before the Net were not spontaneous paragons of epistemic or moral virtue either. Although ''RTFM'' (Read the F***ing Manual) is an internet-age acronym, people have always shown themselves more willing to ask an expert than to try to figure things out for themselves.

Arguably, the human tendency to imitate before innovating or investigating is why cultural evolution is so powerful (Richerson and Boyd, 2006). But without individuals who see and test the limits of what is known, a copycat culture can easily find itself in a cognitive cul de sac. Criticality and innovation are important parts of the engine of culture. However, by making our collective know-how so easily accessible, machines may disguise their ignorance as well as our own. When anyone might consult a computational oracle to get a diagnosis and remedy for whatever appears to ail them, backed up by endorsements from countless users of unknown veracity, the tendency to go along with the crowd is likely to be strengthened. By the apparent certitude of the information computers present or the actions robots take, machines may have insidious effects on our epistemic and moral agency. However, machines open up new possibilities for action at the same time they close down some old ones.

We are not powerless in the face of the globalization of collective ‘‘knowledge’’ that is embodied in our increasingly networked machines, and in the software bots and hardware robots that function with increasing autonomy, i.e. not under direct human control. Among the things we can do is to describe and suggest ways to implement the kinds of virtues, both moral and epistemic, that machines should display. There are many open questions: Should we strive to make machines more like ‘‘wise counselors’’ in the way that they present information? Could they and should they show signs of hesitancy or humility when different answers or different courses of action are available? Can we build machines that are capable of representing and communicating their own limitations? What would be the social and cultural consequences of surrounding ourselves with such machines? Would a kind of faux artificial wisdom be worse, all things considered, than the faux expertise which machines currently instantiate?

We do not know how to answer these questions by philosophical reflection alone. Neither do we think it is feasible to stop the technological developments that are driving us to ask them. Nevertheless, we are not technological determinists. We may not know precisely how philosophical reflection on wisdom will affect the development of artificial intelligence, but we believe that the only viable way forward is to be guided by a combination of technological experimentation and critical reflection on our own limits and the limits of our technologies.

In computational terms, the problem of wisdom shares some features with what computer scientists call ‘‘the frame problem’’. The frame problem is the problem of determining relevance to a task at hand. There is not enough time in the world to consider everything you know for each decision that you make. So how do you frame the decision so as to include only the relevant information? For instance: If planning a trip to Mars, do you need to search your knowledge about camels? The answer may seem to be an obvious no, until you start thinking about issues of water conservation on such a long trip. Maybe something from camel physiology could help.

If you do not know what you do not know, then the problem of determining which facts you should investigate seems insoluble. People use emotional cues to regulate the amount of effort they put into problem solving. When emotional input is lacking, however, the process breaks down. Antonio Damasio (1995) recounts the story of a patient who has brain damage to neural circuitry necessary for processing emotions. The patient’s intelligence is above average, but he reports having very few emotions. He is also incapable of making even simple decisions, for example setting an appointment date. After deliberating fruitlessly for several minutes over two dates offered to him, he simply accepts the one that Damasio picks for him. Without the intervention the patient was off on an interminable search for the things that might affect his decision one way or the other. An emotional predisposition for one option over another as well as an emotional assessment of the significance or insignificance of the decision, and of the relative costs of being wrong, usually prevents people from endlessly mulling over trivial questions.

Intelligence alone does not ensure wisdom. There are individuals with savant syndrome who display amazing intelligence with respect to a very narrow range of mental tasks, psychopaths who are giants of industry, and brilliant yet greedy investment bankers. Computers and robots already outpace us mere mortals in many facets of intelligence. But reflecting on the capabilities necessary for building artificial moral agents forces us to recognize that much more than skill in manipulating symbols is necessary for making wise decisions. Emotions, consciousness, and experience in the world and with other beings inform good choices and actions. Wisdom is a special form of intelligence.

Socrates had an appointment with the executioner, but in the end he did not worry about his own death. He knew that he did not know what would follow, but he also knew that no amount of earthly effort could resolve it. Death was either eternal dreamless sleep, or a chance to converse with his heroes in the afterlife for ever more. In either case, he was happy. After a life of questioning everything and everyone, and urging others to examine their own lives, his emotions told him nothing was to be gained by seeking to learn what could not be learned. That was the wisdom of the philosopher.

References

Damasio, A. (1995), *Descartes' Error: Emotion, Reasoning, and the Human Brain*, Harper Perennial, New York, NY.

Gigerenzer, G. (2008), *Rationality for Mortals: How People Cope with Uncertainty*, Oxford University Press, New York, NY.

Kovalchik, S. (2011), ''How wise was Watson?'', *Significance Magazine*, 18 February, available at: www. significancemagazine.org/details/webexclusive/1019245/How-wise-was-Watson.html (accessed 9 June 20110.

Kurzweil, R. (2005), *The Singularity is Near*, Viking Press, New York, NY. Richerson, P.J. and Boyd, R. (2006), Not by Genes Alone: How Culture Transformed Human Evolution, University of Chicago Press, Chicago, IL.

Vinge, V. (1993), ''The coming technological singularity: how to survive in the post-human era'', paper presented at VISION-21 Symposium, NASA Lewis Research Center and the Ohio Aerospace Institute, March 30-31, available at: www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html (accessed 9 June, 2011).

Wallach, W. and Allen, C. (2009), *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York, NY.

Corresponding author

Wendell Wallach can be contacted at: wendell.wallach@yale.edu