Artificial Intelligence: First and Last Words

> It is desirable to guard against the possibility of exaggerated ideas that might arise as to the powers of the Analytical Engine. In considering any new subject, there is frequently a tendency, first, to *overrate* what we find to be already interesting or remarkable; and, secondly, by a sort of natural reaction, to *undervalue* the true state of the case, when we do discover that our notions have surpassed those that were really tenable. The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we *know how to order it* to perform. It can *follow* analysis; but it has no power of *anticipating* any analytical relations or truths. Its province is to assist us in making *available* what we are already acquainted with.
>
> – Augusta Ada, Lady Lovelace[1]

Introduction

It must be very rare for the first words ever uttered on a controversial subject to be the last word, but Lady Lovelace's comment is just that. Her observation is correct today, just as it was when she made it, and is the basis of an understanding of what computers are and can be, including the possibility of what is now called "Artificial Intelligence"[1]. Her words demand and repay close reading: "[The computer] can do whatever we know how to order it to perform." This means both that it can do *only* what we know how to instruct it to do, and that it can do *all* that we know how to instruct it to do. The "only" part is what most earlier writings on the subject have been concerned with, but the "all" part is at least as important -- it explains (for those open to explanation) why the computer keeps performing feats that seem to show that it's thinking, while critics continue to insist that it is doing no such thing. (Lady Lovelace's views have been the subject of much debate and rebuttal, particularly at the hands of Alan Turing in his celebrated paper "Computing Machinery and Intelligence". His treatment there of her views is examined below; for a fuller analysis of the paper, see [Halpern 1990] and [Halpern 2006].)

As will be argued, the amazing feats made by computers really demonstrate human progress in coming up with algorithms that make the computer do valuable things for us; the computer itself does nothing more than it ever did, which is to do whatever we know how to order it to do -- and we order it to do things by giving it instructions in the form of elementary operations on bits. The more natural-language-like instructions offered by higher-order programming systems seem to show computers understanding something like human language, and thus displaying more intelligence themselves; this delusion is common among those who don't know that such higher-level instructions must first be translated into the computer's built-in machine-language, which offers only operations on groups of bits, before the computer can "understand" and execute them. If you can read Dante only in English translation, you cannot be said to understand Italian.

[1] "Artificial Intelligence" means here "strong AI"; the idea that computers can *think*, in the full sense of that word.

Part I: *Only* what we know how to order it to perform

In fact the computer is the first machine that *can't* think, and its value lies precisely in that disability – it thereby lets *us* think, and preserve our thoughts in executable form. It has so often been touted as the first machine that *does* think, or at least *will* think someday, that this assertion may seem paradoxical or perverse; it is in fact merely unconventional. The computer, as will be made clear, is a most misunderstood device, and the misunderstanding matters greatly.

Why might it seem that computers think?

Speed

Not only does the computer routinely carry out its tasks without error or complaint, it is the perfect servant in another way, too; it is very fast. And mere speed has the power to overawe and mislead human beings, as we all bear witness to every time the house lights dim, and we sink into plush seats, open our bag of popcorn, and wait with mounting excitement as Leo the Lion roars that he believes in art for art's sake. We all know that movies are actually sequences of still images projected onto the screen at a speed that takes advantage of our visual cortex's trick of bridging over the gaps between separate images so as to transform discrete images into continuous action—but our knowing that does not prevent the trick from working. Just so does the speed with which the computer achieves its effects support the notion that it's thinking; if we were to slow it down enough to follow it step by step, we might soon conclude that it was the stupidest device ever invented. That would be just as invalid a conclusion as the opposite, though—the computer is neither intelligent nor stupid, any more than a hammer is.

Personification

Then there is our almost irresistible tendency to personify any device, process, or phenomenon we come into contact with: our houses, cars, ships, hurricanes, nature itself. We invest with personality and purpose just about everything we encounter, the computer among the rest. A recent example of that tendency to personify, and the confusion it can engender: some people seem to be puzzled, even disturbed, by what they perceive as a reversal of left and right (but not top and bottom) in the image they see in a mirror when they stand before it. They believe they see in the mirror a room very like the one they are standing in, with a human figure rather like themselves looking out at them. But that figure, strangely, wears his wristwatch on his right wrist even though we wear it on our left. The solution to this puzzle is, of course, that there is no fellow in the mirror who is facing us from within a room of his own; if there were, we would indeed have the right to be puzzled, but there is only a simple direct projection of our image onto the mirror and back at us, so that the wristwatch is on the left side of the mirror image just as we wear the watch on the left; there is no fellow who has turned to face us, and unaccountably switched his wristwatch to the other side while doing so.

Reinforcing our proclivity for personification, the computer has the special property that its output is often symbolic, even verbal, making it seem especially 'human.' When we encounter computer output that looks like what we humans produce by thinking, we are especially liable to credit the computer with thought, on the illogical grounds that if we are accustomed to finding some good thing in a particular place, it must originate in that place — by which rule of inference there would have to be an orchestra somewhere inside our CD player, a farm in our refrigerator, and live actors up on the movie screen. Urban children are understandably liable to suppose that food originates at the supermarket; AI enthusiasts are less understandably prone to thinking that intelligent activity originates within the computer. Of course, most of the computers we encounter are embedded within other machines and systems, and yield such things as cash, pari-mutuel tickets, and Danish pastries, but we are hardly aware that we are encountering computers in such cases, and draw no conclusions about their intellectual capacity.

Preference for the Positive

And computers have another, even greater thing going for them: the almost universal human preference for the exciting positive rather than the dreary negative. This was noted more than four hundred years ago by Bacon (almost anticipating Karl Popper's rule for distinguishing scientific statements from metaphysical ones): "…it is the peculiar and perpetual error of the human intellect to be more moved and excited by affirmatives than by negatives; whereas it ought properly to hold itself indifferently disposed towards both alike. Indeed, in the establishment of any true axiom, the negative instance is the more forcible of the two."[2]

This natural preference for the exciting is of course enthusiastically shared and gratified by the news media. If it were widely understood that the computer isn't thinking, how could journalists continue to produce those routine pieces titled *Robots—Are They Taking Over?* and *The Computer says "Yes, There is a God—Now!"*, and *Androids: the Next Step in Evolution?*, and all the other familiar scare stories that editors who need to fill up some empty space in their journals regularly commission? These stories implicitly attribute not just intellect to the computer, but *volition* – the computer, if we are to get excited, must *want* to conquer us or replace us or do something equally dramatic. (I forbear to touch on the even more delicate subject of how AI researchers would continue to get funded by various government agencies, foundations, and academic research bodies if it were commonly understood that the computer has neither a mind nor volition.) Nor is the journalism-consuming public an innocent victim in this imposture, but an active co-conspirator. Most of us are far happier reading a sensational, scary-but-not-too-scary story about how machines are threatening or encroaching on us than we are with merely factual accounts of what's going on; nobody loves the spoilsports who go around telling us that there really aren't any man-eating alligators in the sewers. And this is too bad, because the sober truth, in this case, is more interesting than the sensational falsehoods.

Perversity

For reasons that can only be called perverse, the computer seems often to be credited with thinking even when displaying characteristics most unlike those of thinking beings; it is given credit for thinking when in fact it is doing nothing of its own accord, but being absolutely

mindless. (If it does something other than that, we say that there's a bug, and modify the program until it does just what we want it to.) Consider:

- the computer never gets bored (one key indicator of an ability to think is susceptibility to boredom; what thinks – and only what thinks – can be bored.)
- it has a perfect memory
- it executes algorithms faultlessly and at great speed
- it never rebels or becomes contentious
- it never initiates any new action or topic
- it is quite content with endlessly repetitive tasks (what we usually call, tellingly, 'mindless repetition')

Is such an entity thinking – that is, doing what human minds do?

We have gotten the computer to do whatever it does despite our having no complete, or perhaps even rudimentary, theory of intelligence; could we have built a machine that thinks, without thoroughly understanding thinking? Surely, to make machines that think, we must first understand thinking, and to do that, we would have to accomplish something not merely beyond our present powers, but perhaps impossible in principle: we must thoroughly comprehend our own minds, stand outside ourselves as observers, pick ourselves up by our own mental bootstraps. If we can do this, then indeed we will have taken the most momentous step in the evolution of our species—but no one claims we have done this already. And if some day we do, will we need to—will we *want* to—build a machine that reflects this knowledge of what thinking really is? As some wag has noted, all we would achieve by building an artificial thinker is reproduction without sex, when much of the human race seems to want just the opposite.

Chess-playing

A prime example of our tendency to give the computer credit for thinking when it is clearly doing something very different is our crediting it with an ability to play chess, and indeed play it better than any human. When IBM's Big Blue beat Gary Kasparov, many AI champions were ready to write QED; after all, isn't chess-playing ability one of the hallmarks of intellectual power, and if a computer can beat the world champion, mustn't we conclude that it has not merely a mind, but one of towering strength? In fact, the computer does not play chess at all, let alone championship chess. Chess is a game that has evolved over centuries to pose a tough but not utterly discouraging challenge to humans, with regard to specifically human strengths and weaknesses. One human capacity it challenges is ability to concentrate; another is memory; a third is what chess players call *sitzfleisch*—the ability to resist the fatigue of sitting still for hours. The computer knows nothing of any of these. Finally, chess prowess depends on players' ability to recognize general positions that are in some sense "like" ones they've seen before, either over the board or in books. Again, the computer largely sidesteps what is most significant

for humans, and hence for the game, essentially analyzing every position from scratch, and relying on speed to make up for its weakness at *gestalt* pattern recognition.

Surprise!

But the most effective of all the forces making it seem that the computer can think is our occasional surprise at what a computer does—a surprise that puts us in a vulnerable condition, easy pickings for AI enthusiasts: "So, you didn't think the computer could do such-and-such! Now that you see it can, you have to admit that computers can think!" Of course our surprise implies nothing whatever about whether computers can think, but logic is not always trumps on such occasions, and someone whose belief about what the computer can do has just been shown to be spectacularly wrong is in a weak tactical position to resist the claims of the AI champion who was right on the point at issue. So effective is such surprise as a way of extracting concessions from their critics that the Argument from Surprise has become a principal resource of the AI party, trotted out whenever programmers make the computer do something unexpected.

But *surprise* is endlessly fruitful of paradoxes and problems. It names a sensation we cannot even be sure we've experienced; in retrospect, what we once took for surprise may seem no such thing. A stage magician announces that he will make an elephant disappear: it steps behind a curtain; he waves his wand, draws the curtain to show that the elephant has disappeared, and we are duly surprised. But if he had drawn the curtain only to reveal Jumbo still placidly standing there, what would we have experienced? Why, *real* surprise. And if, while we were laughing at the magician's chagrin and our new insight into our own expectations, the elephant suddenly did vanish like a pricked bubble, what we would call our feelings at *that*—and the feelings we'd experienced before? Or again, we are watching a television commercial, and see some terrific vehicle—racecar, airplane, speedboat—piloted with spectacular skill and daring; the vehicle pulls up to a stop; the pilot, a lithe, dynamic figure in pressure suit and helmet, jumps out and strides confidently toward us; and then astonishes us by pulling off that helmet, shaking out an aureole of golden curls, and standing revealed as—a *man*!

To put the question generally, if you've been led to believe that an event will be surprising, are you surprised when it is—or when it isn't? What is it that we expect when we await a surprise? 'Surprise,' for all the air of childlike innocence that clings to it, is, like most human reactions above the level of the knee-jerk, heavily influenced by conventions and expectations of which we are only partly conscious.

To summarize the refutation of the Argument from Surprise: surprise—assuming we can even be sure we've really experienced it—is just the reaction to be expected from humans presented with the result of an elaborate algorithm execution, since such a process is something we ourselves are poor at performing (that's what we invented the computer to do), though good at conceiving and setting in motion (that's why we're the programmers). *Any* process that repeats a sufficient number of times some small, simple, uninteresting step will eventually produce a surprising result**.** The experience of surprise, then, cannot be used as the basis for an argument that the computer is doing something that adds value to, or transcends, what the programmer put into it; what it really is doing—and this is wonderful enough—is yielding full value for everything the programmer *did* put in. This is, to be sure, a result so unprecedented in human experience as

sometimes to seem miraculous, but we must not be so thunderstruck at our success in building a perfectly obedient servant that we take it for our peer, or even our master. Jeeves is marvelous, but is after all a servant; Bertie Wooster is master.

If the Argument from Surprise is discredited, the greatest argument for strong AI falls, but the connection between that thesis and the actual practice of AI workers is obscure. What, if anything, would they do differently if they were convinced of their error? It may well be, in fact, that the claim that machines can or will someday think has no bearing on the day-to-day activities of AI workers, but is simply a morale-building slogan and a bid for funds and prestige. But it is also possible that they have come to believe, as humans so readily do, what they have so often heard themselves say. And as we safely navigated in coastal waters for millennia, despite thinking that the earth was flat, so we may develop interesting and useful programs for many years, despite imagining that we are working with thinking machines. But just as men eventually attempted blue water voyages that came to grief because the navigator thought the earth was flat, so there will be computer projects that will come to grief because managed or attempted by those who think that machines think. Indeed, as will be shown below, at least one serious problem has already been so caused. But even in the absence of such dire consequences, the idea that computers think is damaging to human intellectual coherence

Part II: *All* the things we know how to order it to perform

If it doesn't think, what *does* the computer do?

The place at which to begin to understand the computer, and to appreciate the value of its inability to think, is with the *algorithm*. An algorithm is a complete, closed procedure for accomplishing some well-defined task. It consists of a finite number of simple steps, where "simple" means within the powers of anyone of normal ability and ordinary training, or of anything, such as a computer, that can carry out such simple steps faultlessly. If faithfully executed, an algorithm unfailingly achieves the purpose for which it was created. We have all executed algorithms, at least arithmetic ones—long division, and the multiplication of one multi-digit number by another, are examples. All algorithms are, in principle, executable by humans, but there are a host of potentially valuable algorithms that cannot possibly be carried out, in practice, by unaided humans, because they consist of millions or billions of steps. The computer, however, is perfectly suited to carrying out those algorithms—so much so, in fact, that carrying out those algorithms is the only thing it *can* do. The implications of this fact are not commonly understood, however, and exploring them and making them explicit will be the principal task in what follows.

What makes the computer so ideally suited to the execution of algorithms is that it is the physical embodiment of what was until the latter half of the twentieth century merely a mathematical abstraction: a finite-state machine—a device that can exist in only a fixed number of completely defined and controlled configurations. The simplest and most familiar such device is the ordinary two-way wall switch that we use to turn lights on and off; the computer is made up of billions of such switches. And just as we can always tell which position a light switch is in, and can put it in either position at will, so the computer enables its operator to test the state of any

switch within it; and to make each of them assume whichever state is desired. (The operator can do this manually, at a keyboard or control panel, or programmatically—that is, by causing the machine to execute a stored program.)

And this inability to deviate from the sequence of states chosen for it by the programmer makes the computer the perfect vehicle for any algorithm, and the only possible vehicle for the immensely long algorithms that are the reason for the computer's importance in the modern world. Because the computer cannot depart in the slightest from the sequence of states prescribed for it, it can be used to execute the extremely long algorithms that accomplish tasks that greatly exceed unaided human capacities. And because it is fitted to do this, we begin, paradoxically, to hear claims that the computer is thinking—paradoxical because the special quality of the computer is that it is the first machine we have ever built that clearly does *not* think—that is, that never behaves mysteriously, never appears to have volition, never seems to have a mind of its own. (By "never behaves mysteriously" I mean that it never behaves in a way that, when closely examined, is incomprehensible; for observers who can't be bothered to follow its workings in detail, of course, it is *always* mysterious—but then, everything is mysterious to the inattentive.) In short, the computer seems to be thinking precisely because, being unable to "think for itself," it's the perfect vehicle for and executor of *preserved human* thought – of thought captured and "canned" in the form of programmed algorithms.[3]

While computers don't think, then, they do something far more important, something that is changing human life beyond all imagining: they let us *put up* the fruits of our intelligence, as our great-grandmothers used to put up various foods against the needs of winter. This means that if some genius comes up with an algorithm for accomplishing some task, that accomplishment thereby becomes a permanent possession of the human race, and anyone who can push a button can now accomplish it; for a broad and ever-growing class of human activities permanent and universal mastery is achieved, and the Whig interpretation of history is so far vindicated.

Much of the debate between the critics and the champions of AI has centered around the question of whether human thought can be reduced to mere mechanism, mere organized matter. But the genuinely new element introduced by the computer has nothing to do with mind versus matter, or flesh versus silicon. The great advance it brings is that we now have for the first time a machine that, because it is perfectly obedient and tireless, can carry out for us not only the algorithmic processes we have traditionally carried out in our heads or on paper—generically, "checkbook balancing"—but also a wholly new class of algorithm, never before feasible. That capability means an eventual delegation to computers of all the merely repetitive and mindless activities we find necessary to sustain our lives. But apart from making feasible a vast number of valuable algorithms that were beyond our reach before, the computer does nothing,

We ought to clear 'the computer' out of the way, once and for all, in this and all discussions of AI, since it has no necessary part in them. When we wonder "do machines think?", we are wondering, first, whether some program a machine is executing can be said to be thinking. The computer is merely one possible executor of that program; we could instead organize a regiment of soldiers, each one assigned the role of a bit, to do so. But even the program is extraneous to the fundamental issue; the program is only one of an indefinite number of representations of the underlying algorithm, as there are an indefinite number of sentences that can express a given

proposition. The role of the computer in all this is purely economic; it is the first machine that can execute lengthy algorithms quickly and cheaply enough to make their development economically attractive; it is the algorithm that is ultimately the subject of contention.

The question of "autonomy"

Computers are nowadays credited by some not merely with thinking, but with thinking so independently that we may lose control over them – they are "autonomous", and as such may decide to do things we don't want them to. And this, of course, is especially frightening when the computer is a military robot, armed with weapons that can kill humans.

Back in the last quarter of 2007, articles and books about the ethical implications of military robots appeared with such frequency that it seemed that one could hardly open any serious journal without coming across another warning of the terrible problems posed by these new weapons. Although their frequency was significant, it was not so much their number as the variety of the publications in which such stories appeared that showed that concern over the subject was both deep and broad. In the month of November 2007, almost an entire issue of *Armed Forces Journal* was devoted to the subject of robots as war weapons, with emphasis on how to control them now that they seemed to be on the verge of acting autonomously. And apparently by pure coincidence the November 16, 2007 issue of *Science*, the foremost scientific journal published in America, and perhaps the world, was also devoted to the ethical issues supposedly raised by robots in civilian and military affairs. After 2007 the wave of such stories diminished, and it seemed that journalistic attention had turned to other threats, but in early 2013 we faced a second batch, perhaps because such weapons began to be called "drones," suggesting they are something new, and because the U.S. Government apparently wanted to preserve the option of using them against U. S. citizens. And with this new batch, the specter of robot or drone "autonomy" returned, along with proposals that their use be regulated by an elaborate code and a special judicial apparatus for applying it.[4]

It's not clear whether the fear that some of these pieces raise of robots "taking over" or "getting away from us" actually frightens the writers, or is just the kind of hyperbole that they feel they have to indulge in to catch readers, but there is indeed something frightening in these warnings: they show that there are some widespread misconceptions about the machines we call robots that are going to cause serious problems for us unless corrected very soon. As someone who has been working with computers for many years, and hopes for American economic and military success, I'm alarmed by these misconceptions and want to see the process of correction begin right now.

Note that throughout this essay "robot" and "computer" are treated as essentially synonymous. In the past, a robot was a computer provided with "limbs" – that is, with the means of affecting the world directly and physically rather than by merely providing human beings with information — and limbs that at least roughly resembled human arms and legs. Today a robot need not have limbs that look much like human ones — its appendages are more likely to be tools, weapons, or sensors — but it must seem to act on its own, and hence to raise ethical issues. The discussion that follows is not affected by these distinctions.

Who are Asimov's Robots?

There are two great sources of confusion in typical discussions of robots and their supposed ethical issues. The first is that we seem to have a rich literature that we can draw upon for insights and guidance: science fiction. We have whole libraries of books and magazines, going back at least as far as Mary Shelley's *Frankenstein*, dealing with the creation by human beings of creatures whose behavior they cannot fully control, and pondering the question of who should be held responsible for that behavior. If literary critics and scholars in the humanities were consulted about the problem, it would not be at all surprising if they referred to this literature to see what earlier thinkers had come up with. What is surprising — highly surprising — is that supposedly hard-headed scientists and military officers are doing the same thing, under the impression that this literature was dealing, however fancifully, with the very issues we are now called upon to deal with in earnest. The first step toward sanity in the matter, then, will be to expose that misunderstanding.

The acknowledged dean of the modern imaginative literature on robots was Isaac Asimov (c. 1920—1992). For over half a century, the astonishingly prolific Asimov wrote science-fiction stories and novels about future civilizations in which humanoid robots play a central role. He made himself so much the proprietor of this subgenre of literature that many of the premises about man-robot relationships first introduced in his writings have been adopted by other writers[5]. In particular, his "Three Laws of Robotics," first announced in one of the earliest of his robot stories, are now virtually in the public domain, and during his lifetime Asimov had to assert his authorial rights to them repeatedly and vigorously. These laws, wired unalterably into the "Positronic brains" that are the seat of the robots' intelligence, are:

> **First Law**: A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
>
> **Second Law**: A robot must obey orders given it by a human being except where such orders would conflict with the First Law.
>
> **Third Law**: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

.

The Asimovian Robot gets taken more seriously than Asimov intended

The difficulty of seeing how the Three Laws are to be applied in specific circumstances, let alone embodied in electronic circuitry, is so great that it's clear that they are simply literary conventions, allowing readers to enjoy the stories while ignoring the many practical problems that would otherwise concern them. The Laws make the robots ordinary decent people, trying to

be good in the face of various problems that are difficult enough to be interesting, yet not so terrible as to leave them sunk in despair.

Asimov's robots, then, are really people in light armor; but except as a footnote to literary history, does it matter? It does, and very much so: the Asimovian robot is what almost everyone envisions when the topic of robots and how to control them comes up, as it is now doing with some urgency. The special issue of *Science* has already been noted, but the November 2007 issue of *Armed Forces Journal (AFJ)*, also devoted chiefly to robots – specifically, to robots as weapons of war – is one that deserves more attention, because it may very well reflect national military policy. Although the two periodicals could hardly be more different in ethos, they are united in holding a view of robots that is false and dangerous – and in large part the legacy of Isaac Asimov. (Asimov himself is very little to blame for this; he was quick to disclaim any expertise in actual computers or man-computer interaction; it is his millions of readers who seem unable, when robots are the topic, to distinguish fiction from real life.)

The gist of the message that Asimov, however unwittingly, has passed on not only to the general public but even to many technically sophisticated people – computer scientists, engineers, mathematicians -- is that robots, if not already behaving with a large measure of autonomy, will be doing so very soon. And this, if true, makes it a matter of some urgency to answer such questions as whether to allow robots to take human life – a question usually put in the form, do we want a computer or a human being to decide whether to kill someone? – or whom to blame if a robot does something bad, and even to wonder what rights a robot may have, since it apparently has some degree of free will.

The guest editorial in *Science*, for example, includes these comments:

> As we make robots more intelligent and autonomous, and eventually endow them with the independent capability to kill people, surely we need to consider how to govern their behavior and how much freedom to accord them – so-call roboethics. Science fiction dealt with this prospect decades ago; governments are wrestling with it today.

And later:

> Again, science fiction may be our guide as we sort out what laws, if any, to impose on robots and as we explore whether biological and artificial beings can share this world as equals.

The idea that Asimov's Three Laws (they are what Sawyer mainly has in mind, although he also mentions an even more simplistic formulation by the science-fiction writer Jack Williamson), which were useful to their author simply because they gave his imagination the space it needed, should help governments decide matters of life and death, is a chilling one. (And his reader has to hope, for the sake of Sawyer's personal welfare, that either he is unmarried or that his wife never sees the concluding sentences of his editorial, in which he speaks of "…one viable future: a man marrying a robot woman, and living, as one day all humans and robots might, happily ever after. I, for one, look forward to that time.")

But the notion that Asimovian Laws and the conventions of science-fiction in general can be enlisted as a guide in formulating real-world rules is just one of the two great misconceptions that are bedeviling modern political and military thinking about how to use robots. The other great misconception entertained by supposedly realistic policy makers is that we have succeeded, or are about to succeed, in creating robots with some degree of autonomy, and this error is due to the popular attitude toward the research effort that has been going on since the middle of the twentieth century under the name of "the strong Artificial Intelligence project".

Artificial Intelligence as corrupter of human intelligence

The subject of Artificial Intelligence (AI) has generated a vast literature, if nothing else, and to trace its history in any detail would require a multi-volume work, not an essay. (I have offered elsewhere much fuller treatments of the subject than is possible here; readers who want to see the claims of AI discussed thoroughly are referred to my book[6] and a more recent article[7] in which I treat AI from its origin in Turing's Imitation Game up though the Loebner Prize Competitions and Searle's Chinese Room thought experiment). But even the highly abbreviated treatment that I can offer here should be enough to indicate why AI has had a corrupting effect on popular thought, and on what is supposed to be responsible and serious policymaking.

The modern AI program began with the publication in October 1950 of Alan Turing's landmark essay "Computing Machinery and Intelligence" in the British philosophical journal *Mind.* As every schoolboy should know, Turing at least implied there that a computer should be judged intelligent if it could carry on a reasonably extended conversation with a human interrogator with such fluency that he couldn't be sure whether he was talking to another person or to a computer. In making this proposal, Turing gave us another way of dichotomizing the human race. Some people find this proposal to be common-sensible and refreshingly free of the metaphysical claptrap about the "mind" that takes that entity as somehow mystical and ineffable; others find it too silly to be taken seriously. I am one of the latter group.

One reason for rejecting Turing's criterion for intelligence is that it is more a measure of the interrogator's feelings and expectations than of anything scientific – that is, rigorous and repeatable; it is much more a test of human credulity than a controllable experiment. Another reason for rejecting the Turing Test (TT), as it is now known (Turing himself modestly called it the Imitation Game), is that having been a computer programmer myself, and an observer of the AI project almost since the year Turing began it, I have noted that all the considerable effort that has been put into realizing Turing's thought experiment has resulted in exactly nothing — nothing, that is, that constitutes progress toward Turing's vision. A great many byproducts have come out of projects labeled AI, some of them useful and ingenious, but one thing that has never come out of them is AI itself, particularly passing the Turing Test. Since the idea of machines that really think is exciting, journalists are forever issuing overheated reports about supposed AI breakthroughs. These are invariably the achievement by programmers and engineers of applications of the computer to new areas, such as driving an unmanned vehicle through rugged terrain; a potentially very useful feat, but one that has nothing to do with the TT or with endowing a computer with intelligence.

Some AI advocates, realizing that their work, however fruitful in other ways, is making no progress toward achieving what Turing wanted, have tried to move the goalposts. They claim that passing the TT is no longer the goal of AI, or even that it never really was — but they fail to provide an acceptable alternative. What becomes the *de facto* criterion when the TT is abandoned is the degree of astonishment felt by observers at each new product emerging from the AI workshop; AI champions of this school contend that if computers achieve something highly unexpected, we have to grant them intelligence. But this proposed test, like the TT, is essentially a measure of what an observer happens to value and expect rather than of some new general capability of the computer, and is to be rejected for the same reason. As noted earlier, some people were highly impressed by the victory of a computer programmed to play chess over the human world champion; those who know something of computers, and of chess, were not impressed at all — they have always known that a computer with enough speed and a big enough database of chess positions would beat any human player, just as a bulldozer can move more dirt than a man with a pick and shovel.

Another way in which AI enthusiasts have tried to turn their disappointing results into victories is to claim that computers may not have achieved human intelligence, but have achieved a different kind of intelligence. This is trying to win an argument by changing the definition of a key term in mid-discourse; allowed this tactic, one can prove anything at all. These rather desperate moves have caused even emeritus Professor Marvin Minsky of MIT, who with John McCarthy of Stanford founded and named the Artificial Intelligence project in the United States, to say "AI has been brain-dead since the 1970s…. For each different kind of problem, the construction of expert systems had to start all over again, because they didn't accumulate common-sense knowledge….Graduate students are wasting three years of their lives soldering and repairing robots, instead of making them smart. It's really shocking."[8]

But the reader of American newspapers and popular magazines has been bombarded for so long by stories about breakthroughs in AI, many of them promising or announcing great things, some warning him that computers might soon take his job away from him or even take over the whole world, that many have a vague notion that there are, or soon will be, machines that are intelligent in the full sense; machines that think and act autonomously. In most cases this delusion is harmless; when entertained by military planners and other policy makers it is potentially greatly harmful.

The truth is that we have not given any computer, whether in robotic or any other form, the slightest degree of autonomy, nor is there the slightest reason to believe that we could do this if we wished to. Computers sometimes evince behavior that their programmers do not expect; as noted earlier, when that unexpected behavior is unwelcome, we say that there's a bug in the program, and modify the program until it yields the wanted results. But if the unexpected behavior is welcome, we are too often tempted to say "It's thinking! It came up with better results than I had any right to expect; we have Artificial Intelligence!" This is a parade that needs to be rained on quickly and forcefully, before it leads us further into error; when a program yields results different from what its programmer expected – whether better or worse -- what it means is that the programmer doesn't thoroughly understand his program; that is, cannot predict

exactly what it will do, even though it is executing a purely determinate program of his own creation. There is nothing strange about this; if we humans were good at working out in our own minds the full consequences of a very lengthy algorithm, we wouldn't need computers, we'd *be* computers. But whether the programmer can predict its behavior or not, the computer *never* does anything other than what its program makes it do; if we don't fully understand our programs, that doesn't mean that the computer is thinking for itself -- it has no self, and what has no self does not think — all it means is that we have not fully realized the consequences of the program we loaded into it.[9]

We should indeed be embarrassed at this silly error of mistaking our own shortcomings, in memory capacity and in ability to follow very long chains of reasoning, for corroboration of the claims of Artificial Intelligence and computer autonomy, but there is the more serious danger that we will sincerely come to feel that misdeeds committed by robots are the fault of the robots themselves, rather than of their programmers and users. The *AFJ* editor's introduction to the issue on military uses of robots says, "The next ethical minefield is where the intelligent machine, not the man, makes the kill decision." But no existing or presently conceivable robot will ever make such a decision; whether a computer kills a man or dispenses a candy bar, it will do so because its program, whether or not the programmer intended it, has caused it to do so when certain conditions are met. The programmer may be long gone; it may be impossible even to identify him; the action he has inadvertently caused the robot to take may be one he would abhor – but if anyone is responsible for the robot's action, it will be either the programmer or the commander who, despite knowing that robots have no common sense, dispatches one to deal with a situation for which common sense is required; in no case will it be the robot.

As a consequence, the question of responsibility for a robot's actions is a bogus one. Our society, in both peace and war, regularly sets in motion processes that are quite robot-free, but already so complex that it is rarely possible to attach blame or credit to any one person for their outcomes. We recognize this when we refer off-handedly to the "Law of Unintended Consequences" — a law that began to manifest itself long before robots entered the picture. Events on a battlefield are particularly hard to trace back to any one actor or even unit; historians still argue about the role of Blücher at Waterloo: would Wellington have won the day even without the Prussians, or was he saved from disaster only by their arrival? No one knows, and no one ever will. The entry of robots on what is already a confused and tumultuous scene adds no further problems; if anything, they clarify one corner of the picture: we will, at least usually, be able to tell *what* a robot did, if not whether it was good or bad. This is more than we are usually able to determine about human battlefield actions, where confusion, misunderstandings, and accidents — the well-known "fog of war" — make it hard to determine even the simplest facts about what went on during a battle.

The designer of a robotic system can of course put any number of layers of processing between the system's initial detection of a situation and the actions that it will finally perform as a result, and if he puts in enough such layers, it may well become impossible for him to foresee just how the system will deal with some particular situation. This uncertainty may cause observers — and sometimes, amazingly, even the designer himself — to think that the system is deciding the

issue; but they are wrong if they do. The system may well be executing the sequence of steps built into it much faster than anyone can follow, but that does not mean that it is exercising its own judgment, for it has none. Yes, it may *seem* so to many, just as a series of still photographs, projected at the right speed, will convince our eyes that they're seeing continuous motion – but we have learned to believe our reason and experience, and not our lying eyes, even while enjoying the movie. And we will undoubtedly continue to personify robots and other computerized systems as we personify everything; it's been a habit of the human race from our beginnings, and does no harm provided we are aware of it and are prepared to correct it when necessary, as it is here.

The whole "should a man or a robot decide?" debate, then, is wrongly framed, and asking the wrong question makes it impossible to get a useful answer. *The decision will always be made by a man; a robot can't decide anything; it can only do what it's been programmed to do.* The robot may be so programmed as to perform actions that the programmer cannot, even in principle, predict — but that will only mean that its program includes a random-number generator, either explicitly or, as when buried in the data used in training a Machine Learning program, implicitly. The only question is whether the man making the decision will be the designer/programmer of the robot, who may have made the decision ten years before the robot is called on to act, or a user working with the robot at the moment of its use, and directing it to do one thing or another at that time. That is indeed a very serious question, and much depends on reaching the right answer for any given application, but getting the question itself wrong guarantees the answer will be worse than wrong, it will be irrelevant. The real question a robotic system designer has to face is how much authority to assume himself, and how much to put into the hands of the ultimate user of the robot – not the robot itself.

The advantage of having the designer/programmer make the decision is that it will be made calmly and deliberately by someone working in peace, with advice and input from others available as he needs them. He will have time for testing and revising and debugging, and won't suffer from panic or frantic haste. The decision he programs into the system, provided that he has not overlooked some critical factor, is likely to be the best available. The disadvantage is that if he *has* overlooked some critical factor, whether because he was less than perfect as a designer or because a new and unforeseen element has entered the picture since he created the program, the robot may do something disastrous.

The advantage of having the user make the decision at the time of use is that he will be able, at least in principle, to take into account the current details of the situation in which the robot is being applied, and to make whatever changes may be necessary to adapt the robot to the actual situation. The disadvantage is that the user will be acting on the spur of the moment, perhaps in panic mode, and without the benefit of anyone else's inputs.

Which is the right way to go? There is no general right way; it depends on the specifics of the situation. There will be some systems so complex that it would be ridiculous to expect any user to be able, in the heat of battle, to correct an oversight made by a designer/programmer – more likely a team of designer/programmers – who were working in far better conditions, probably during peacetime. An ABM (Anti-Ballistic Missile) defense system is an example of a

computerized system whose complexity is such that no user can be expected to make last-minute improvements to it at its moment of use in anger; for better or worse, whatever decisions its designers built into it when they did their job is what we will have to live – or die -- with.

At the other end of the spectrum of possibilities, a UAV looking for a terrorist leader, and carrying a small bomb or rocket to deal with him when he's found, is an example of a system that the user should be enabled to control right up to the last minute; the decisions to be made are so dependent on new and unforeseeable details that any ordinarily competent user will be able to make them more accurately than the designers who created the system he's using. Most robotic applications, I think, would fall at some intermediate point along the spectrum whose end points were just sketched, and a lot of careful thought should be spent on them in order to arrive at an optimum balance for each of them. But nothing of value at all can be accomplished if we get the question wrong; as Bacon said, Truth emerges more readily from error than from confusion. And so we return to where we began; what we have to do is soberly think through the best way to handle each type of situation, knowing that our robots will do exactly what we have told them to (not necessarily what we *want* them to), and leave the dreams of thoughtful, loving, and — above all, autonomous -- robots to Asimov and his successors.

Part III: Turing's treatment of Lady Lovelace's views

Turing's treatment of Lady Lovelace's views, particularly of her refusal to grant originality to the computer, exhibits his typical mixture of arrogance and carelessness. About 10 pages of his paper are devoted to criticizing nine objections, selected by himself, to his own thesis. Of these, the two titled "Lady Lovelace's Objection" and "The Argument from Consciousness" are the only ones that deal with actual objections made by named persons; all the rest are attributed to unnamed and possibly imaginary persons, or to no one at all. It is not clear that, except for the two just named, any of the objections had been expressed by anyone; it may well be that they were dreamt up by Turing to give him a chance to promote his own views. While rejecting these supposed objections with the condescension of a scientist dealing with the confusions of the laity, he himself, amazingly, says of the objection that he calls "The Argument from Extra-Sensory Perception", that "This argument is to my mind quite a strong one", and that of the four components of E.S.P. – telepathy, clairvoyance, precognition, and psycho-kinesis – the statistical support for at least the first of these "is overwhelming." He offers no evidence for this statement.

In his attempt to refute or evade Lady Lovelace's total rejection of the idea of computer "thinking", Turing seizes on the absence of the word "only" from her statement "It can do whatever we *know how to order it* to perform," suggesting that its omission leaves open the possibility that she did not mean utterly to deny the possibility of computer originality (Turing1950, p. 459, n.1). Her thought as she wrote this sentence, then, would be "In truth, I think that the computer *can* do more than just follow our instructions, but I will express that thought simply by refraining from using the word 'only' before "whatever we know", and I will neither here nor anywhere else mention its wonderful capacity for originality." If it seems to the judicious reader unlikely that this was her thinking, there is an alternative interpretation: that she omitted "only" because it was too obvious to need saying after the immediately preceding

sentence “The Analytical Engine has no pretentions whatever to *originate* anything.” She was seeking to characterize the computer fully, both its limitations and its capabilities, and would hardly have omitted any mention whatever of its capacity for originality if she thought it possessed such a thing. And in fact Turing knows perfectly well that she meant “*only* whatever we know”; when paraphrasing her views on another page (Turing 1950, p. 454), he says “Let us return for a moment to Lady Lovelace’s objection, which stated that the machine can only do what we tell it to do.”

Hollings 2018, p. 82, states:

> Alan Turing disagreed [with Lady Lovelace's statement]. In a famous paper on 'Computing Machinery and intelligence' he challenged what he called 'Lady Lovelace's objection'. He suggested that the machine could be 'ordered' to be original, by programming it to produce unpredictable answers.

I have been unable to find in Turing's paper any passage clearly saying what Hollings et al. tell us he said, but if he did indeed say that, it was a claim remarkable for incoherence even in that poorly reasoned paper.[10] The computer is to be "ordered" to be original? And by being "programmed" to produce "unpredictable answers"? When we call someone's utterance "original", we ordinarily mean not only that it is new – at least to us -- but that it has some merit; at the very least, that it is coherent and intelligible. If it is not even intelligible, we do not call it original, we call it nonsense. Turing (assuming that he really made this strange claim) gives no assurance that the "unpredictable answers" that would be the result of our orders to the computer would be even intelligible, let alone of some value. When the computer produces output that is "original" by such slack criteria, we groan, and start looking for the bug.

What Turing definitely did, though, was to ask that the computer be assigned credit for whatever unexpected discoveries could be made in its output. In a radio address he delivered on BBC Radio on 15 May 1951, called “Can Digital Computers Think?”, he said:

> If we give the machine a programme which results in its doing something interesting which we had not anticipated I should be inclined to say that the machine *had* originated something, rather than to claim that its behavior was implicit in the programme, and therefore that the originality lies entirely with us. (Turing’s emphasis)

Against this inclination of Turing’s I would raise three points: first, why need we assign credit for the unexpected interest of the results to *anyone*? If we find a diamond in the street, do we try to decide who gets the credit for our find? Second, consider our treatment of unexpected scientific results: when scientists perform an experiment, they may acquire valuable knowledge they had not explicitly sought; if they do, to whom is credit given? Whoever gets it, it’s certainly not the apparatus they used in the experiment, although it may be the humans who designed that apparatus. Third, and most important, how did we discover that the program has produced something original or interesting? It wasn’t the computer that discovered that, it was the programmer or his client or some other human observer who decided that the output has value.

So if we need to assign credit for that discovery, I suggest that it's the discerning investigator who identified the jewel in the muckheap who is entitled to it.

And the amount of credit to be awarded for computer-aided "breakthroughs" may be too skimpy to be worth quarreling over. When, for example, the computer was used by Appel and Haken to solve the classic four-color mapping problem, there was a curious apathy on the part of mathematicians; they acknowledged some good work on the part of the investigators in transforming the problem into one that could be answered by exhaustion of a finite number of possibilities, but the result itself was felt to be without savor or excitement. What the episode pointed up was the unimportance of merely getting an answer to the problem—not even cartographers care much whether four colors are, in principle, sufficient to color a map—as contrasted to the growth of insight that would normally be gained en route to that answer. In giving us the answer without such deeper understanding, by magic, the computer left us feeling that we were both cheated and cheating, as if we'd gotten to the top of Everest by helicopter. Andrew Gleason, late professor of mathematics at Harvard, remarked:

> I detect a monstrous amount of boredom among mathematicians in response to this proof. Generally, mathematicians are interested less in which theorems are true, and more in why they're true. This solution relies on a known style of approach—the one first used by Kempe and extended by Birkhoff and others—and blasts the final answer out with a computer. I know very little more about the four-color problem now than I did before. I've talked to Haken about this, and he says little to show that he has any more insight into it, either.

What has happened in the case of the four-color problem is prophetic of what will happen with AI triumphs generally. Programs called "AI" will achieve great usefulness, even—in some fields—indispensability, but they will not be hailed as their fond authors wish them to be; that prize will forever elude them. For as they seem to play better chess than Capablanca and prove theorems that would have defeated Gauss, it will become equally well understood that they aren't doing any of these things in interesting ways—ways that deepen human understanding of the problems being dealt with. The computer will not really be playing chess or proving theorems, but merely executing the appropriate algorithm quickly enough to seem miraculous, at least to naive observers. The results produced may well be invaluable, but they will come at too high a price if they corrupt our understanding—as they will, unless the machine that produces them is understood to be just another example of how we spin off specialized devices to handle problems we have in principle solved, so as to free ourselves for the mental play we call thinking.

In the absence of any generally accepted alternative goal or criterion, it is practically impossible to say what is and what is not AI. Any new software that comes out of an institution with "AI" in its title, or that is developed by a graduate student whose thesis advisor teaches a course on "AI", is usually called AI when it first appears—and who can contradict such a claim? But these "exciting developments" and "breakthroughs" are always demoted to plain old applications when their novelty has worn off. The result, as AI workers frequently complain, is that the strong AI thesis fails to benefit from anything they do—all their triumphs are soon thought of as just more software. "It's a crazy position to be in," laments Martha Pollack, a professor at the Artificial

Intelligence Laboratory at the University of Michigan and executive editor of the *Journal of Artificial Intelligence Research*. "As soon as we solve a problem, instead of looking at the solution as AI, we come to view it as just another computer system," she told *Wired News*.[11] One sympathizes with Professor Pollack, but there is no help for her; she is no doubt familiar with Clarke's Law: "Any sufficiently advanced technology is indistinguishable from magic," but she may not have appreciated that "is indistinguishable from" is a symmetric relation, just as readily yielding Halpern's Corollary: "Any sufficiently explained magic is indistinguishable from technology."[12]

Turing and the Turing Test: their Role in Modern AI Thought

The weaknesses of the TT, only a few of which have been explored here, have made both the man and the Test highly problematic for AI enthusiasts, who want to enlist Turing as their spiritual father and philosophic patron. While they have programmed the computer to do things that might have astonished even him, today's programmers cannot do what he believed they would do—they cannot pass his test. And so, the relationship of the AI community to Turing is much like that of adolescents to their parents: abject dependence accompanied by embarrassed repudiation. For AI workers, to be able to present themselves as "Turing's Men" is invaluable; his status is that of a von Neumann, Fermi, or Feynman, just one step below that of immortals like Newton and Einstein. He is the one undoubted genius whose name is associated with the AI project (although his status as a genius is not based on work in AI). The highest award given by the Association for Computing Machinery is the Turing Award, and his concept of the computer as an instantiation of what we now call the Turing Machine is fundamental to all theoretical computer science. When members of the AI community need some illustrious forebear to lend dignity to their position, Turing's name is regularly invoked, and his paper referred to as if holy writ. But when the specifics of that paper are brought up, and critics ask why the Test has not yet been successfully performed, he is brushed aside as an early and rather unsophisticated enthusiast. His ideas, we are then told, are no longer the foundation of AI work, and his paper may safely be relegated to the shelf where unread classics gather dust, even while we are asked to pay its author the profoundest respect. Turing's is a name to conjure with, and that is just what most AI workers do with it.

At the start of this essay I referred to the puzzlement some people experience at the apparent fact that the figure we see in the mirror seems to have switched left to right, and I offered the explanation that the figure we seem to see there is illusory; there is no 'man in the mirror'. I offer that explanation again, this time to explain the illusion some suffer about computer thinking: there is no one in the computer, hence no one doing any thinking. We humans see evidence of thought in the running computer, as well we might, but we fail to recognize the thinking party as ourselves. And the observation of Lady Lovelace remains the foundation of real understanding of the computer and AI.

A note on the new AI: "Machine Learning" (ML) or "neural networking"

Since approximately 2015 there has been a resurgence of excitement over AI, and not only in programming circles. Industrial and military organizations throughout the world are all trying to

take advantage of a new technical development in programming that bears a variety of names – see the section title just above -- but which is founded on the idea that giving computers enough data and a very general goal like "Find the common element!" enables them to produce useful information for us without our knowing exactly how they're doing it.

A commonly used example of such an application is that of getting the computer to recognize images of things of interest – cats, say -- without being told how to recognize them. This is done by (1) having the computer so programmed scan thousands of pictures of cats of many different kinds, (2) having it extract from these pictures some features it finds to be common to them, and saving these common elements as a tentative solution, and then (3) feeding it as a test another batch of pictures, some of cats and some of other things, and asking it to distinguish the cat-pictures in this test batch from the non-cat pictures. Experience to date is that the computer can do that, if not perfectly, at least with significantly greater success than chance would account for. What is exciting to law-enforcement and military officials about such a process is that, assuming it enjoys continued success, it could be applied to finding terrorists in pictures of crowds, or defects in manufactured articles, or targets in pictures of terrain.

The programs that do this are modeled roughly on our knowledge of the structure of the human brain as multiple layers of interconnected neurons, although the importance of this rough analogy is not clear. The programmer begins by assigning neutral default weights to each layer of neurons in his program. If when tested they recognize cats better than chance would have it, or better than previous results, they get "rewarded" – their weight is increased. if they fail to do so, they are "punished" by having their weight decreased. It is hoped that this process, analogous to evolution by natural selection, will produce an ever-improving cat-recognition program, and it has had enough success to date to cause a great deal of excitement and investment of resources.

What such a program is looking for when it decides whether or not a given picture shows a cat, we don't know; it could be that the program has noted that in all the samples given it of cat pictures, a certain pixel in the upper right corner of each picture was "on", and that's what it looks for rather than anything we think of as feline. At present the program's idea of the significant feature, whatever it is, is known to itself alone, and that's where the champions of AI find their opening: maybe in learning to recognize cats without being told what to look for, the "computer is thinking!" But unfortunately for the enthusiast, some investigators are already trying to determine just what it is that the new AI programs are doing in their internal levels. When they succeed, we will be back in the situation so lamented by Professor (now President) Pollack, that of seeing an AI "breakthrough" or "major advance" as just more software – which is what it will be.

I have provided here only a tiny and superficial picture of the new ML development; I have not, for one thing, bothered to distinguish between the supervised and the unsupervised varieties. I have neither the space nor the qualifications to provide a full and authoritative picture of it, but for present purposes the description provided will do. *The only aspect of this new technology with which I am concerned here is the effect it is having on the perennial question of whether computers can or will think, and the answer to that is clear.* The practical benefits of this latest development are as yet unknown; it may turn out to be a great step forward, or just another overblown laboratory phenomenon, like the Fuzzy Logic craze of several years ago, or the

Formal Methods of proving programs correct of a decade or so before that. But in the meantime, it is causing a new burst of excitement over AI, and a recrudescence of the old fantasies about machine thinking and the new ones about autonomy – and in doing so, it is a step backward.

But does a misunderstanding of computers and AI matter?

As I've said, I regard the intellectual incoherence of believing that computers think, or can be autonomous, is sufficient by itself to warrant forceful correction, but for those unconcerned with such coherence, there are practical dangers that may be more impressive.

On April 25, 1984, in a hearing before the Subcommittee on Arms Control of the United States Senate Committee on Foreign Relations, a heated dispute broke out between several Senators, particularly Sen. Paul E. Tsongas (D-Mass) and Sen. Joseph R. Biden Jr. (D-Del), and some officials of the Executive Branch, particularly Robert S. Cooper, the Director of the Defense Advanced Research Projects Agency (DARPA), and the President's science advisor, George Keyworth, about reliance on computers for taking retaliatory action in case of a nuclear attack on this country.

The Associated Press story states that the controversy began when the Executive Branch witnesses acknowledged that a space-based laser system designed to cripple Soviet long-range missiles in their 'boost' phase would have to be triggered on extraordinarily short notice. To strike the boosters before they deployed their warheads in space would require action so fast that it might preclude a decision being made in the White House—and might even necessitate a decision by *computer*, the panel said [emphasis supplied].

> At that, Sen. Tsongas exploded:
> "Perhaps we should run R2-D2 for President in the 1990s. At least he'd be on line all the time."
>
> "Has anyone told the President that he's out of the decision-making process?" Tsongas demanded.
>
> "I certainly haven't," Keyworth answered.
>
> Sen. Biden pressed the issue over whether an error might provoke the Soviets to launch a real attack. "Let's assume the President himself were to make a mistake..." he said.
>
> "Why?" interrupted Cooper.
> "We might have the technology so he couldn't make a mistake."
>
> "OK," said Biden. "You've convinced me. You've convinced me that I don't want you running this program."

The key point here is that the two parties, although diametrically opposed on the issue of whether computer involvement in a missile launching system was a good thing or not, agreed that such involvement might be described as "letting the computer decide" whether or not missiles were to be launched, in contrast to "letting the President decide".

Neither party seemed to realize that the real issue was the utterly different one of *how* the President's decision was to be implemented, not whether it was his or the computer's decision that was to prevail; that what is really in question is whether the President, in the event of nuclear attack, will seek to issue orders by word of mouth, generated spontaneously and on the spur of the moment, or by activating a computer-based system into which he has previously had his orders programmed. The Director of DARPA accepted as readily as did the Senators that the introduction of the computer into the system would be the introduction of an independent intelligence that might override the judgement of the duly elected political authorities; he differed only in considering that such insubordination might be a desirable thing.

These political and military leaders of the United States were conducting their debate, and formulating national policy, on the basis of this nonsensical assumption because hundreds of journalists and popularizers of science have succeeded in convincing too many of the computer laity that computers think, or almost think, or are about to think, or can even think better than humans can. And this notion has been encouraged and exploited by many computer scientists, who know or should know better. Many apparently feel that if the general public has a wildly exaggerated idea of what computer scientists have accomplished in this direction, so much the better for their chances when funds are disbursed, and authority granted. The point at which AI ideology affects important real-world matters has, then, already been reached; here it is corrupting a debate among decision-making parties on one of the most urgent issues of the day.

And as these words are written, the newspapers tell us that Joseph Biden is currently the leader in the contest to become the Democratic Party's candidate for U. S. President in 2020.

References

Author	Date	Title	Publisher
Copeland, B. Jack	2004	*The Essential Turing*	Oxford University Press
Essinger, James	2014	*Ada's Algorithm*	Melville House
Halpern, Mark	1990	"Turing's Test and the Ideology of Artificial Intelligence," in *Binding Time: Six Studies in Programing Technology & Milieu*	Ablex Publishing Corporation
----------	2006	"The Trouble with the Turing Test"	*New Atlantis*, Winter, 42-63
Hollings, Christopher; Ursula Martin, and Adrian Rice	2018	*Ada Lovelace: The Making of a Computer Scientist*	Bodleian Library U. of Oxford
Turing, Alan Matheson	1950	"Computing Machinery and Intelligence"	*Mind* 59, 433-460

Endnotes

[1] Augusta Ada King, Lady Lovelace (1815-1852), known informally as 'Ada Lovelace', was a colleague of Charles Babbage, the inventor of the Analytical Engine. When she translated Luigi Menabrea's memoir describing that computer, she added some notes of her own. The epigraph here is part of her Note G; the emphases are hers.

Among the misunderstandings that surround her life and work is the claim that she was the first computer programmer. The basis of this claim is the table she included in Note G in which she recorded each successive state that would be assumed by the Analytical Engine as it computed Bernoulli numbers. But her table is not a program -- a program is a sequence of *instructions* that bring about the successive states of a computer as it is executed, not a record of the *states* themselves. It is true, though, that the table is logically equivalent to a program: the program can be inferred from it. What she produced is as if a chess reporter should record a game not by specifying the moves made by the players, but instead depicting the successive states of the chessboard as the game proceeded, move by move. So while it is strictly inaccurate to say that Lady Lovelace wrote the first program, she did something just as noteworthy – indeed, she did all that could be done at the time. It would have been impossible for her to record the program itself in a readable form, since there was then no programming language; the instructions to the Analytical Engine were in the form of punched cards, not written representations to be entered via a keyboard and printed out on demand. (See Hollings 2018, page 79, for details).

[2] Aphorism XLVI, Book One, *Novum Organon*

[3] It would be well for us to abandon once and for all the very name *computer;* it has the merit of reflecting the machine's historical origins, but also the serious fault of misleading us about the machine's real nature. The computer should be called, formally, an Algorithm Executor—for short, an Alex (but *not* a Smart Alex). Mathematicians and logicians should recognize it as the physically realized calculus, or model, of standard predicate logic, such that to every step recognized by that logic there corresponds a physical step in the machine, and the execution of that step in the machine yields the representation of the corresponding step in the system of logic—as some logician has neatly put it, "the sum of the representations is the representation of the sum".

[4] See Scott Patterson, "Letting the Machines Decide," *Wall Street Journal* (July 14, 2010); Bill Keller, "Smart Drones," *New York Times* (March 17, 2013), p. 7; Robert H. Latiff & Patrick J. McCloskey, "With Drone Warfare, America Approaches the Robo-Rubicon," *Wall Street Journal* (March 15, 2013), p. A13; and Kenneth Roth, "What Rules Should Govern US Drone Attacks?," *New York Review of Books* (April 4, 2013), pp. 16, 18.

The Roth piece, in finding humans responsible for whatever the "drones" do, is more sensible than the great majority of those on the subject, which are concerned about the supposed autonomy of these devices, but it is so stringently legalistic that it winds up being just as unrealistic as the others. Roth wants every concept, every doctrine, as rigorously defined as terms in mathematics — what exactly, he demands to know, is a "combatant," as distinguished from a "civilian"? When exactly is a threat "imminent"? How much harm must a military power accept before deciding that it has arrived at "the last resort" where war is justified? It is unlikely that a committee of legal experts could come up with universally acceptable definitions of these terms anytime this century; the only thing one could be sure of is that while they were conferring, the terrorists would be killing. But Roth wants more: not only are the requirements of that semi-mythical entity called "international law" to be satisfied, but so are those of another and even more nebulous one called "humanitarian law" — something I'd not heard of before, but expect to hear more of in future.

The quality of the Keller piece is adequately indicated by a subhead reading "Coming soon: weapons that have minds of their own."

[5] And not only writers; Dr Joseph Engleberger, once widely considered the *doyen* of industrial robotics, and chairman of HelpMate Robotics, has said "I want to make a robot that is in the image of the principles set out by my mentor, Isaac Asimov." Quoted in *Red Herring*, August 2000, page 226.

[6] *Binding Time: Six Studies in Programming Technology & Milieu* (Ablex, 1990).

[7] "The Trouble with the Turing Test," *The New Atlantis* (Winter 2006,) pp. 42-63

[8] quoted in Mark Baard, "AI Founder Blasts Modern Research," *Wired News* (May 13, 2003), pages 1-3, at pages 1 and 2.

[9] I dismiss the claims of the Artificial Intelligence (AI) champions rather unceremoniously here; readers who want to see those claims discussed more thoroughly are referred to my "The Trouble with the Turing Test," in *The New Atlantis* (Winter 2006,) pp. 42-63, and to the relevant chapters of my *Binding Time: Six Studies in Programming Technology & Milieu* (Ablex, 1990). Between the two, I deal with AI from its origin in Turing's Imitation Game up though the Loebner Prize Competitions and Searle's Chinese Room thought experiment.

[10] In emails to Hollings on 5/12/2019 and Martin 5/21/2019 I asked for an exact citation, but have had no replies.

[11] Professor Pollack's remark is quoted in the same story in which Marvin Minsky's disappointment with the direction that AI research was taking is reported (see Note 8, above). Her distress over the unfair reception that AI achievements get may be somewhat assuaged by her elevation to the presidency of Cornell University.

[12] This relationship has been noted in a very different context: Keith Thomas writes in his *Religion and the Decline of Magic* (Penguin ed., 1973), p. 799, "...what is not recognized by any particular observer as a true 'science' is deemed 'magic' and vice versa."
.