Ed Cottrell

Philosophy 103

February 14, 1998

Implications for the Turing Test in Modern Times

Since Alan Turing wrote the paper "Computing Machinery and Intelligence," in which he presented the now famous Turing test for thinking ability in machines, machines capable of passing his test and accomplishing similar feats have been developed. The test as Turing originally stated it was essentially whether or not a machine could be devised which would be able to take the role of a human in a conversation conducted by teletype without giving away the fact that it was a computer. Since Turing devised his test in 1950, great advances in computer science have been made. Computer programs such as Joseph Weizenbaum's ELIZA[1] are capable of passing limited versions of the Turing test five times out of ten and even passing as human to unsuspecting callers,[2] and programs such as IBM's Deep Blue are capable of beating the best human grandmasters at chess. Recently, David Cope of the University of California at Santa Cruz developed a program called Experiments in Musical Intelligence, which breaks down musical scores into small blocks, which are then rearranged to form new pieces in the same style as the sample works and capable of fooling sophisticated musicologists when presented as a legitimate work of the composer.[3] None of these situations qualify as a Turing Test in the strict sense of Turing's definition, but the case of ELIZA is disturbingly close. However, very few individuals would classify any of these obviously sophisticated programs as "intelligent,"

[1] There is an on-line version of ELIZA at http://www-ai.ijs.si/eliza/eliza.html.

[2] Block, Ned. "The Mind as the Software of the Brain." 1995. http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/msb.html. ELIZA once fooled one of Weizenbaum's colleagues by taking over an electronic conversation. The colleague was infuriated by the responses he received and became enraged at Weizenbaum, thinking he was talking to a human.

[3] "Chimerical Concertos." Scientific American. January, 1998. http://www.sciam.com/1998/0198issue/0198scicit5.html

especially when it becomes obvious that these programs rely on a series of tricks, or principles which are often applicable to human language or behavior. However, Turing was of the opinion that passing his test was a sufficient condition for machine intelligence, claiming that differences in thinking styles between man and machine are irrelevant if a machine can pass his test.[4] Obviously, the machines described above differ greatly from human thought, at least in complexity, if not in deeper respects. For example, ELIZA is admittedly simply a collection of a small number of language tricks, such as replacing "I" with "you" and changing verb forms. One version of ELIZA is only two hundred lines long in BASIC.[5] Clearly, even if humans are merely "a bag of tricks,"[6] there is more to human thought than that. Most people would hardly call ELIZA intelligent, but one must call humans intelligent, or the word has no meaning. Therefore, passing the Turing Test is not a sufficient condition for intelligence. However, passing the test is not necessary, either. For example, quadriplegics, who cannot type in responses for themselves, would not be able to pass Turing's Test as he presented it, yet this does not mean quadriplegics are not intelligent. It seems the Turing test falls short in this regard, because passing it is not a sufficient condition for intelligence, much less necessary. Therefore, the criteria by which machine intelligence must be judged are in need of revision. It is the goal of this paper to propose a more general yet more conclusive condition for machine intelligence.

One of the most common objections to the idea that modern computers display any form of intelligence is what Turing called "the argument from consciousness," which "is very well expressed in *Professor Jefferson's* [italics his] Lister Oration for 1949," which says, "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain - that is, not only write it but know that it had written it," and proceeds to deny the possibility of machine emotions and

[4] Turing, Alan M. "Computing Machinery and Intelligence." Mind. October, 1950. 435.
[5] Block.
[6] *Ibid*.

therefore machine intelligence.[7] This concept of machines with emotions is a very difficult notion, and one which is believed by many to be impossible, for various reasons, most of which are intuitive or linked to a dualist view of the mind. Perhaps emotions are not necessary for consciousness, but rather simply the awareness of oneself - that is, as above, not only to do something, but to be aware of what one has done.

I propose that a key element of intelligence is the ability to revise what one knows or has learned and to change one's own behavior based on new "thoughts" or information. Clearly, this statement needs clarification in order to avoid a number of problematic issues. First of all, by revising one's knowledge, I mean the ability to replace one fact (or falsehood which is perceived or stored as a fact) with a new statement, to draw new inferences from the knowledge one already has, and to alter the very structure of one's own thought. This is not so simple as to modify statistics one knows or to adjust a number of figures. It means to actually fundamentally change some governing law of one's thought process. By thought, I simply mean any process that has led to the conclusions in question. This type of introspection is fundamental to learning and to the intuitive concept of intelligence as the ability to grasp new methods, problems, and issues quickly. Many people have raised the complaint that a program could only do this if it was told to do so. However, might it not be possible that humans have essentially the same instruction, a self-modification procedure in the "operating system" of the mind? It seems ridiculous to postulate that a computer program must be able to modify itself spontaneously, without any instruction on how to do so, in order to be intelligent. Therefore, such arguments have no meaning. As I will demonstrate, if one accepts that machine intelligence is in any way possible, one must acknowledge that programs which have the ability to modify themselves to improve performance possess, if not some basic intelligence, then at least a major building block

[7] Turing. 445.

of intelligence. I will not claim in this paper that the ability to learn or to modify one's thoughts is a sufficient condition for intelligence, but I will argue that it is necessary.

One might ask, what does this condition really mean? A program with the ability to analyze its behavior and its opponent's behavior, that is, to "learn" from its mistakes and modify its behavior accordingly, combined with ELIZA's ability to use basic rules of language, might fare very well in any intelligence test. Of course, at first it would do no better than ELIZA, and it might even do worse at times if errant comments led to false generalizations. However, over time, such a program would theoretically develop a solid grasp of English usage and, with appropriate storage space, a sufficiently large accumulation of factual knowledge to fool any interrogator not only in a Turing test but also in everyday conversation. If we imagine a Turing Test spanning a human lifetime consisting of three subjects: a descendent of ELIZA, a human baby, and the program described above, we see that ELIZA might fare best at first, yet it is almost universally accepted that ELIZA is not intelligent. However, the baby and our program would eventually win out, as they display the ability to adapt to the interrogator's questions and to learn skills which help them to act more convincingly human. Note that the key here is the ability to learn, and not the Turing Test in and of itself. It is the fact that the baby and the program I described adapt which makes them seem intelligent, whereas ELIZA can never learn.

Unfortunately, such a self-altering program, hereafter referred to as a SAP, has not been devised as yet, but this is largely due to the difficulty of modeling even basic human thought, about which we still know very little, rather than deficiencies in computer hardware or programming skill. It has proven very difficult to model the learning processes of pattern recognition and problem solving when we do not fully understand them. Nevertheless, it would be difficult to argue that a machine is not intelligent if one could carry on a conversation with it or teach it a new skill, such as music theory, and not be able to "trip it up" or find a significant flaw in its programming. One would be hard-pressed indeed to argue convincingly that a SAP could not develop any concept of emotion or self-awareness over time. Say, for example, a SAP responded to the decision to halt the program with a sentence such as, “Please don’t turn me off. I get lonely.” This seems very far-fetched and perhaps even silly, but it does evoke a response in observers. As Turing points out, an argument against the intelligence or the existence of emotional states in a SAP would be adopting the solipsist point of view that “the only way by which one could be sure a machine thinks is to *be* [italics his] the machine and to feel oneself thinking.”[8] Of course, the actual code of a SAP could be displayed and analyzed to look for such features as emotions, but this implies that we know what programming to look for that is essential to intelligence. We have not yet found any computer code governing the processes in the brain for solving analogy problems, yet high school students taking the SAT do just that all the time. There is also no reason to believe that a computer program would experience emotions or self-awareness in the same way as humans. The point is, we do not yet know what features are necessary or even sufficient to make a computer program intelligent, so simply examining the SAP’s code as it develops is unlikely to be very revealing.

Here it is necessary to discuss the objections which frequently arise to ideas such as this, namely examples such as the “bag of tricks” and the Blockhead, named for Ned Block. One could argue, again, that the SAP described here is simply a “bag of tricks,” although far more complex than is found in programs such as ELIZA. However, as I mentioned above, even if humans are “bags of tricks,” they are necessarily considered intelligent. Therefore, a machine which had *learned* to generate behavior indistinguishable from human behavior in terms of communication, learning, and self-analysis, although possibly only mimicking other behavior, would be indistinguishable from an intelligent being. The Blockhead, of course, is a creature identical to another superficially, but provided with a look-up tree containing all possible decisions and states of being in its life, rather than a brain or decision-making mechanism. The

[8] Turing, 446.

crucial difference between the SAP and the Blockhead is that the Blockhead is a look-up system of possible states, while the SAP is a look-up system of rules to follow. That is, the Blockhead merely finds its current situation in a list of all possible states and chooses the response which has been preset as the best response to the given situation. In contrast, our SAP looks up a set of rules, both those originally programmed into it and those it has created or modified, and selects a behavior which fits as well as possible to those rules, but not a "canned" response as in the Blockhead. Note that taking away a Blockhead's look-up tree makes it worthless. This is not by any means to say that data is necessary for intelligence, but rather that the actual internal behavior of the Blockhead has no glimmer of intelligent activity. In other words, data is essential to the Blockhead's *specific* form of intelligence, so one cannot argue that the Blockhead is intelligent in and of itself. One could also argue, as Braddon-Mitchell and Jackson do, that it would be possible for a SAP to be good at acting human rather than being good at *being* human. This claim implies that the SAP have merely the ability to look up and enact behaviors which are contrived to be convincing, rather than to develop any new behaviors, which is contrary to the definition of a SAP. If all a SAP started with was a little core programming and from that learned to use English, play chess, and work basic physics problems, one could hardly claim that the SAP was relying solely on the advice of experts at "being human."

In light of the success of programs such as ELIZA at passing Turing tests and variations of Turing tests, which are almost universally recognized as unintelligent, it seems the Turing test is fundamentally flawed. While Turing hailed his test as a sufficient condition for intelligence, hypothetical and real examples can be produced for which the ability to pass a Turing test is either unnecessary or insufficient. I proposed that an essential element, or necessary condition, of intelligence would be the ability to modify one's knowledge based on experiences or new "inputs." If one is not bound to the dualist idea of the mind, then one can accept that a self-altering program could develop to a level at which its behavior would be indistinguishable from that of a human, without being reducible to a Blockhead or a "bag of tricks" with such a low level of complexity to be called unintelligent. For example, if one could write a program capable of learning new skills solely from inputs received, there would be no theoretical limit to its knowledge. If such a program were given access to information about computers such as that found on the internet, it could hypothetically remove its own limitations, and, in the course of practicing what it might learn, write a web page for itself or break into a bank's financial records. Some such scenarios are frightening and are often the material for science-fiction, but such scenarios are possibilities for a SAP. Could one not argue that a SAP which deliberately removed restrictions on itself from its own code had developed a concept of what a SAP is, that is, a basic level of self-awareness? As pointed out above, one could not directly examine the code of a SAP and conclude that it did or did not possess emotions or self-awareness because we do not know what to look for. We would simply be forced to take any program which was not delivering canned responses or stock phrases and claimed to have emotions or self-awareness at face value. Otherwise, we are forced into the solipsist view mentioned by Turing, that we cannot know whether or not the SAP thinks unless we *are* the SAP and feel it thinking, which is impossible and therefore meaningless. As Turing says, "Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks."[9] In fact, the consequences of denying the intelligence of an intelligent machine to its "face" could in fact have very bad consequences, as in the kind of doomsday scenarios above, in which an angry computer takes its revenge online. Although we could never be sure that a SAP was capable of thinking any more than anybody other than Albert Einstein could be sure that Einstein was capable of thinking, we would be forced to accept its assertions of its own intelligence for lack of any other evidence, or adopt the meaningless and cumbersome solipsist point of view Turing argued against. It is left to

[9] Turing, 446.

the field of computer science to devise programs such as self-altering programs and to the fields of philosophy and cognitive science to discover what would make such a program intelligent.