2010

Moral Agency and Advancements in Artificial Intelligence

Arielle L. Zuckerberg
Claremont McKenna College

CLAREMONT McKENNA COLLEGE

MORAL AGENCY AND ADVANCEMENTS IN ARTIFICIAL INTELLIGENCE

SUBMITTED TO

PROFESSOR AMY KIND

AND

DEAN GREGORY HESS

BY

ARIELLE ZUCKERBERG

FOR

SENIOR THESIS

FALL 2010

NOVEMBER 29, 2010

ACKNOWLEDGEMENTS

I would not have completed this project if I had not had the support of my family and friends. First, I would like to thank Professor Amy Kind for helping me narrow down my topic and realize what it means to have a strong work ethic. I would also like to thank Harry Schmidt, my brilliant brother-in-law, for coaching me through the final stretch. Finally, I would like to thank my wonderful parents for believing in me.

TABLE OF CONTENTS

ABSTRACT

The main trends underlying advancements in AI technology are increased autonomy, increased functionality, higher level social interaction and integration, and, many suggest, hard-wired ethical codes of conduct. These trends, along with the promises of engineers and scientists, give us ample reason to evaluate the agency of future AI machines.

INTRODUCTION

The formally unimaginable world in which planes and cars steer themselves, surgeries are performed by machines, and babies learn about social interaction by observing robots, has become our reality. Researchers at Google have built a fully automatic car that has clocked in over 1,000 miles on the road without human intervention, and one even drove itself down Lombard Street in San Francisco.[1] McGill University Health Center in Canada recently hosted the world's first automated anesthesia and robotic surgery.[2] Unsurprisingly, people's attitudes toward their mechanical helpers seem to be changing as well. Researchers at the University of Washington just published a study concluding that babies are more likely to treat social robots as sentient beings than inanimate objects. [3] Emerging technology offers even more sophistication and autonomy in artificially intelligent agents. NASA, for example, hopes to send a robot to the moon within three years; its forerunner has already been sent into

[1] Sebastian Thrun, "What We're Driving at," The Official Google Blog, October 9, 2010, http://googleblog.blogspot.com/2010/10/what-were-driving-at.html (accessed November 2, 2010).

[2] Emma Woollacott, "First All-robotic Surgery and Anesthesia Performed," TG Daily, October 20, 2010, http://www.tgdaily.com/general-sciences-features/52099-first-all-robotic-surgery-and-anesthesia-performed (accessed November 2, 2010).

[3] Molly McElroy, "I Want to See What You See: Babies Treat 'social Robots' as Sentient Beings," University of Washington News, October 14, 2010, http://uwnews.org/article.asp?articleID=60848 (accessed November 2, 2010).

orbit to help maintain the space station.[4] Armed, autonomous, robotic soldiers that can distinguish between comrades and enemies are already in development.[5] Autism researchers see a bright future in humanoid therapists helping autistic patients forge stronger social bonds.[6] The extensive competency of present and future AI agents ensures that we will rely heavily on their functionalities and come to expect a certain level of productive output from them.

Interestingly, the notion of machine intelligence emerged as a result of research about the nature of human intelligence. Norbert Wiener's notable studies about feedback loops led him to theorize that intelligent behavior in humans resulted from feedback loop mechanisms, and that machines could potentially simulate these mechanisms.[7] Wiener's research had a strong influence on the early development of AI and inspired people to reevaluate their conceptions of human intelligence. The idea that human intelligence itself consists mostly, if not entirely, of computational processes has captured the interest of engineers who seek to stretch the boundaries of current technology and blur the lines between human and machine intelligence. This computational model provides a scientific paradigm that effectively connects neuroscience and psychology with AI. It seems that we have a pervasive urge to recreate ourselves.

---

[4] Kenneth Chang, "NASA's Quest to Send a Robot to the Moon," *The New York Times*, November 2, 2010, Science section, New York edition, http://www.nytimes.com/2010/11/02/science/space/02robot.html (accessed November 5, 2010).

[5] Jason Palmer, "Call for Debate on Killer Robots," BBC News, August 3, 2009, http://news.bbc.co.uk/2/hi/technology/8182003.stm (accessed November 5, 2010).

[6] Gregory Mone, "The New Face of Autism Therapy," *Popular Science*, June 1, 2010, http://www.popsci.com/science/article/2010-05/humanoid-robots-are-new-therapists (accessed November 5, 2010).

[7] Norbert Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*, 2nd ed. (Cambridge: MIT Press, 1965).

Today, artificially intelligent agents reside outside of our moral realm due to their limited capabilities, but as machines become more sophisticated and autonomous, we must decide how we should think about the question of what we owe our machines. Given that there will be machines that can pass the Turing Test, what conditions are required in order for these agents to be part of a moral community?

My goal is to establish some ethical framework for how we should approach the question of what moral privileges we owe our machines. In chapter one, I offer a description of the Turing Test, explain what it tests, and discuss its relationship with what is required for an agent to receive consideration for moral status. I summarize and refute the most significant objections to the Church-Turing thesis. In chapter two, I engage in a deeper exploration of the concept of moral agency and begin to frame the question of what we owe our AI machines.

CHAPTER ONE

I. THE TURING TEST

Six decades ago, when the Information Age was in its infancy, Alan Turing dared to approach the question, "Can computers think?" But it was not long before he realized that this question was riddled with philosophical ambiguities. Even our notion of what it is to "think" remains a topic of debate. In his 1950 paper, *Computing Machinery and Intelligence*, Turing introduces what he calls the "Imitation Game" and proposes a new set of questions to replace our original, "Can computers think?"

The imitation game involves three players: one human interrogator, one human player, and one machine player. The interrogator asks the players questions through an input terminal in order to determine which player is human and which is machine. Both players try to act in a way such that the interrogator perceives him (or it) to be human. A machine has effectively "passed" the Turing Test if the interrogator cannot reliably tell the machine from the human. Turing therefore proposes that we replace the question, "Can computers think," with the question, "Are there imaginable digital computers which would do well in the imitation game?"[1]

Turing narrows down the list of eligible machine participants by specifying that the machine player must be a “digital computer.” A toaster, for example, would not be

[1] Alan M. Turing, "Computing Machinery and Intelligence," Mind 59, No. 236 (1950): 433-60, doi: 10.1093/mind/LIX.236.433 (accessed September 30, 2010).

eligible to participate in the Turing Test, because it is not a digital computer; it is merely an appliance.[2] Digital computers, Turing explains, "are intended to carry out any operations which could be done by a human computer."[3] The "human computer" is the way he refers to the human capacity for thought (or more vaguely, the "mind"). He notes, "The human computer is supposed to be following fixed rules; he has no authority to deviate from them in any detail. We may suppose that these rules are supplied in a book, which is altered whenever he is put on to a new job. He has also an unlimited supply of paper on which he does his calculations."[4] From this point, Turing maps out how the functions and parts of a digital computer correspond to those of a human computer. A digital computer, he claims, can be regarded as consisting of three parts:[5]

(i) Store.

(ii) Executive unit.

(iii) Control.

The digital computer's store corresponds to the human computer's calculation scrap paper, rule book, and memory (for example, when the scrap paper is not needed to perform the calculation). The digital computer's executive unit is the part that performs the calculation step-by-step, much like how the human computer performs the calculation. The individual operations in each step will differ greatly from machine to

---

[2] I do not doubt that there will be highly sophisticated toasters in the future, but in this example, I am referring to the typical, present-day, household toaster oven.

[3] Turing, "Computing Machinery and Intelligence," 436.

[4] Ibid.

[5] Ibid., 437.

machine, just as some human computers are more mathematically inclined than other human computers. Lastly, the role of the digital computer's control is to make sure that the rules in the rule book (stored in its program memory) are obeyed.

Turing's detailed outline of these correspondences show that not only have digital computers been modeled after human computers, but they can also "mimic the actions of a human computer very closely."[6] Furthermore, the process by which a digital computer is programmed to perform a certain task is similar to how a human computer learns to carry out a new operation. Learning in humans consists largely of reinforcing specific patterns of behavior and discouraging others, both purely mechanical processes.

The thrust of Turing's essay is against the exceptionalism concerning the human capacity for thought that was pervasive in his day, and especially evident in the objections to his thesis. His goal is not to identify whatever the *sine qua non* of the human mind is and attempt to ascertain whether a computer can emulate it. Instead, he calls into question our own assumptions about what "intelligence" is.

What makes the Turing Test so brilliant is that it black-boxes the entire discussion about the nature of human intelligence. Whatever that intelligence is or however it works, if another human being communicates with an AI machine and cannot tell the difference between it and an actual human, then for all intent and purposes, Turing argues, it is exhibiting "human"-type intelligence. In order for a machine to conceivably pass the Turing Test, it must implement problem-solving and linguistic skills, as well as inductive and deductive reasoning, not to mention substantial background and contextual

[6] Ibid., 438.

knowledge. Turing is not concerned with the definitions of consciousness or spontaneity and how they relate to human intelligence. He believes that these are misleading lines of argument because they try to draw attention to the differences in the way the intelligence is packaged rather than address anything meaningful about its underlying nature.

In 1950, Turing predicted that by the turn of the millennium, "it will be possible to program computers with a storage capacity of about 10^9, to make them play the imitation game so well that the average interrogator will not have more than a 70 percent chance of making the right identification after five minutes of questioning."[7] Today, the average laptop computer has a storage capacity of about 10^{10} machine words, and many desktop computers have 10^{11} (a terabyte). But despite Turing's underestimation of our developments in storage capacity, we have not yet engineered a machine that stands a chance of passing the Turing Test.

II. OBJECTIONS TO TURING'S THESIS

In his paper, Turing preemptively refutes nine objections to the notion that machines can think.The most significant of these in terms of ethics are "The Argument for Consciousness" and "Lady Lovelace's Objection." I summarize both of these objections and explain why they fail to undermine the validity of Turing's thesis. In

[7] Ibid., 442.

addition, I describe John Searle's well-known "Chinese Room Argument," which I rebut more thoroughly in the next chapter.

Lady Lovelace's Objection

Lady Lovelace's objection is that analytical engines (the largely theoretical mechanical computers of her own time, a century before Turing) do not have the ability to originate ideas (unlike humans, who can generate original thoughts and ideas). Turing cites a quotation from her memoir: "'The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform' (her italics)."[8] Turing says, in agreement withDouglas Hartree's position in *Calculating Instruments and Machines* (1949), that perhaps the machines to which Lady Lovelace had access did not display this ability, but that observation has no bearing on the possibility for future machines to have the ability to originate anything:

> It will be noticed that [Hartree] does not assert that the machines in question had not got the property, but rather that the evidence available to Lady Lovelace did not encourage her to believe that they had it. It is quite possible that the machines in question had in a sense got this property. For suppose that some discrete-state machine has the property. The Analytical Engine was a universal digital computer, so that, if its storage capacity and speed were adequate, it could by suitable programming be made to mimic the machine in question.[9]

Her objection, however, does raise important questions about originality and the concept of learning. Turing entertains the thought that originality could be a misconception, and what we consider to be "original work" could just be the result of

[8] Ibid., 450.

[9] Ibid.

following and applying principles that we have been taught. He wonders, "Who can be certain that 'original work' that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles?"[10] This thought is a subset of the notion that we do not create new things, but rather we discover them.[11] Turing explains that if this notion is true, then perhaps a better variant of Lady Lovelace's objection is that a machine can never be spontaneous or "take us by surprise." The reasoning behind this objection is that if we are diligent in our calculations of a machine's expected output given a known input, then the actual output will always be the expected output and nothing else. Only a miscalculation on the part of the human will lead to "surprising" output from the machine, but the machine of course cannot receive credit in this case because the output is a result of human miscalculation rather than the machine itself. Turing replies that the view that machines cannot give rise to surprises is due to unfounded assumptions about the way in which machines are constrained by their hardware and software versus the way in which humans are constrained by biology and physiology. Those who accept this variant of Lady Lovelace's objection generally assume that constraints on machines prohibit spontaneity whereas constraints on humans do not. But if the behavior of an artificially intelligent machine is deterministic, why would human behavior be any different? It is very possible that human behavior is deterministic as well, for it is unclear that the ways in which we think and

---

[10] Ibid.

[11] The "discovery" principle goes back to Plato's philosophical treatise *Meno,* where the philosopher purports to show that learning is not a creation of something new but rather a remembrance of something old.

learn are different from simply running a program. Perhaps in human to human interaction, the instances in which we believe we have been taken by surprise are indeed just instances of miscalculation on our part, and like the machine, the other human cannot receive credit for the "surprising" output.

The Consciousness Objection

The consciousness objection is that machines cannot think because they cannot feel emotions or perceive sensory input as being pleasurable or painful.[12] It denies the validity of the imitation game by appealing to the view that consciousness equates with *qualia* (feeling states) rather than cognitive processes. Turing admits that there is a sort of mystery about consciousness, but the mysteriousness is not exclusive to the consciousness of machines. He reminds us of the classic mind-body problem that arises from the "consciousness as *qualia*" viewpoint, which is that the only way one can be certain that another person thinks is to be that person and experience oneself thinking:

> According to the most extreme form of this view the only way by which one could be sure that machine thinks is to be the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise according to this view the only way to know that a man thinks is to be that particular man.[13]

Therefore, those who accept the consciousness objection are either solipsistic, or simply making an unfounded assumption that humans can feel but machines cannot. In regards to the latter, it is not uncommon for humans to not know with whom they are

[12] Ibid, 445.

[13] Ibid, 446.

communicating over the Internet. This occurs most often in chat rooms, forums, and comments sections. In most cases, we assume that we are talking to other humans, despite the fact that many of us are aware of the possibility that we could be talking to computer generated responses.[14] Several websites implement chatterbots to assist users, and they are often programmed to speak like humans to give the illusion that a human behind the computer is dedicated to helping users with their problems. Even ELIZA, the "computer therapist" who first appeared in the 1960s, was frequently mistaken for a human, despite the limited technological resources available to her programmer, Joseph Weizenbaum.[15]

Oftentimes, we are willing and prepared to "accept our invisible interlocutor as a (normal, ordinary) human with human thinking capacities."[16] If those who accept the consciousness argument believe that some being is human, they would also find it justifiable to believe that the being has human thinking capabilities. But if they realize that they are mistaken, and that the being is actually a machine, they would immediately change their minds about the justifiability of their belief in the being's human thinking capabilities. Turing finds it rather naïve to base one's view of whether or not a belief is justifiable on a factor that is so dependent on assumption, and that one's beliefs must shift

---

[14] For example, receiving emails, text messages, and instant messages consisting of computer-generated spam that masquerades as human-generated content in order to gain the trust of readers is extremely common. Unfortunately, many people are tricked into downloading harmful viruses or giving away their credit card information as a result of this spam.

[15] Joseph Weizenbaum, "ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM* 9, No. 1 (1966): 36, http://web.archive.org/web/20071026055950/http://www.fas.harvard.edu/~lib51/files/classics-eliza1966.html (accessed November 24, 2010).

[16] William J. Rapaport, "How to Pass a Turing test," *Journal of Logic, Language and Information* 9 (2000): 469.

in accordance with the truth of that assumption. Moreover, it seems ignorant to doubt a being's cognitive capabilities solely based on the being's provenance (biological versus mechanical).

The Chinese Room Argument

Turing's dismissal of this objection is problematic, since he does not address the deeper issue, namely that his imitation game is not a test of intelligence, but rather a test of the ability to *simulate* intelligent behavior. This concern is more clearly delineated in John Searle's "Chinese Room" argument, which addresses the nature of human intelligence head-on. Searle argues that behavioral properties of intelligence alone are not sufficient grounds for determining whether or not an agent is intelligent, and likens the distinction between machine intelligence and human intelligence to that of syntax and semantics. His goal is to prove that passing the Turing Test does not adequately indicate intelligence, and he draws a distinction between what he calls "strong AI" and "weak AI." In cases of strong AI, Searle explains, "the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states."[17] On the other hand, in cases of weak AI, the computer merely simulates thought and cannot be said to truly understand or have cognitive states. He explains, "According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and

[17] John R. Searle, "Minds, brains and programs," *Behavioral and Brain Sciences* 3 (1980): 417-57, doi: 10.1017/S0140525X00005756 (accessed September 30, 2010).

test hypotheses in a more rigorous and precise fashion."[18] In Searle's view, the Turing Test attempts to test for intelligence in accordance with strong AI, but it inevitably fails and instead achieves accordance with weak AI.

Searle's *Minds, Brains, and Programs* (1980) introduces the "Chinese Room" thought experiment to present his objection Turing's thesis and disprove the strong AI hypothesis. He asks the reader to imagine that she is "locked in a room, and given a large batch of Chinese writing" plus a "second batch of Chinese script together with a set of rules for correlating the second batch with the first batch." The rules are in English, and enable the reader to "correlate one set of formal symbols with another set of formal symbols".[19] He uses the word 'formal' to mean that the reader "can identify the symbols entirely by their shapes." The reader is then given a third batch of Chinese symbols and another set of instructions that allow her "to correlate elements of this third batch with elements of the first two batches," and her final task is "to give back certain sorts of Chinese symbols with certain sorts of shapes in response." Searle adds, that unbeknownst to the reader, the first batch is called "a script" by those who give it to her. Likewise, the second batch is called "a story," the third is called "questions," and the set of rules is called "the program." These names, however, have no bearing on how the reader performs the task. He then asks the reader to imagine that she has become so adept at following the instructions and performing the task, that "from the point of view of somebody outside the room," her responses to the third batch of Chinese symbols are

[18] Ibid.

[19] Ibid.

"absolutely indistinguishable from [answers to the questions] of native Chinese speakers."[20] Even though she speaks and understands no Chinese and she has no idea that the third batch of Chinese symbols are questions or that her responses are answers to these questions, people consistently believe that the reader is a native Chinese speaker based on her responses. Now suppose that the above process is repeated, except with English writing instead of Chinese. Again, the reader follows the instructions and performs the task so well that her responses are naturally "indistinguishable from those of other native English speakers" (which makes sense, because she is a native English speaker). Searle points out that from a third-party perspective, the reader's answers to the Chinese questions and her English questions are "equally good." But the processes by which she formulates her responses in the Chinese case are fundamentally different from those in the English case. Specifically, in the Chinese case, she "produces the answers by manipulating uninterpreted formal symbols," whereas in the English case, she understands that she is receiving questions and giving answers, and the symbols have semantic meaning that has little to do with their shapes. Searle argues that in the Chinese case, the reader "simply behave[s] like a computer" since all she does is "perform computational operations on formally specified elements." She is essentially an "instantiation of the computer program."[21] She receives input that is meaningless to her and produces output likewise.

---

[20] Ibid.

[21] Ibid.

In Searle's view, a native Chinese speaker understands Chinese. The reader, however, alone in the "Chinese Room," does not, because her mere instantiation of the computer program is not a sufficient condition for intentionality.[22] Yet she has fooled everyone into believing that she is fluent in Chinese. Searle believes that this is precisely what occurs when a machine passes the Turing Test. Like the reader, the machine fools the interrogator with its presumed natural language aptitude, but in reality, the machine does not understand what it is saying or what is being said to it. Therefore, Searle claims, it is possible for an entity to pass the Turing Test without being able to think, and the strong AI hypothesis fails.

III. REPLIES TO THE CHINESE ROOM ARGUMENT

There are several responses to Searle's argument that seek to reinforce the validity of Turing's thesis. In this chapter, I analyze the most popular replies, namely, the Systems Reply and the Robot Reply, as well as the Reply to Searle's Argument from Semantics and the Reply to Searle's Notion of Intentionality.

The Systems Reply

The most common response to Searle's Chinese Room argument is known as the Systems Reply, which says that it is true that the reader, alone in the Chinese Room, does

[22] Ibid.

not understand Chinese, but that is because she is just a part of a larger system, and that larger system could indeed understand Chinese.[23] The larger system includes the reader, the set of instructions, and the scrap paper (memory) she uses to help aid the process of responding. She plays the role of the central processing unit (CPU), or, as Turing calls it in his essay, the executive unit. All parts of the system are required in order for it to understand Chinese.

Searle frames his description of the Systems Reply such that it "simply begs the question by insisting without argument that the system must understand Chinese."[24] However, the Systems Reply should be understood as an attack on Searle's logic. It subtly proves that his argument against the possibility of the system understanding Chinese is not logically valid. The proposition that the reader's formal symbol manipulation does not enable the reader to understand Chinese does not entail the proposition that the reader's formal symbol manipulation does not enable the Chinese Room as a whole to understand Chinese.[25]

This reply proves to be quite problematic for Searle's argument. Clearly, the reader would not be able to convince anybody that she were a native Chinese speaker if she lacked the comprehensive set of instructions in English. It is the combination of the reader plus the instructions that fools everybody, not the reader alone. The instructions are paramount to the system's success in appropriately responding to specific questions in

---

[23] Ibid.

[24] Ibid.

[25] Jack B. Copeland, "The chinese room from a logical point of view," in *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence*, ed. John M. Preston and Michael A. Bishop (Oxford: Oxford University Press, 2003), 110.

Chinese. However, in the conclusion of his thought experiment, Searle regards the reader as the sole entity participating in the Turing Test. He then fails to see how that would be similar to a machine's CPU taking the Turing Test without that machine's memory store or instruction set. A CPU alone does not qualify as a digital computer, and therefore, Searle has not proved that an entity could conceivably pass the Turing Test without being able to think, and Turing's thesis stands.

Searle's response to the Systems Reply is that a man could theoretically memorize the entire instruction manual and do all of the calculations in his head, thus internalizing the entire Chinese-understanding system, but the formal symbols would still have no meaning to him. He could *be* the system without understanding a word of Chinese.[26]

This counterargument is most eloquently shot down by Jack Copeland in his paper, *The Chinese Room from a Logical Point of View*. Copeland argues that the man who internalizes the system is still just the CPU, or the implementor, despite the fact that the internalized system does fully understand Chinese. He likens this to the mechanisms in the brain that solve equations in a manner so precise that they enable us "to catch cricket balls and other moving objects," yet we do not understand how or why these equations work.[27]

The Systems Reply is but one response to Searle's argument that seeks to preserve the validity of the Turing Test as an indicator for intelligence, and it succeeds by

---

[26] Searle, "Minds, brains and programs."

[27] Copeland, "The chinese room from a logical point of view," 112.

uncovering a level-of-description fallacy in Searle's claims. Other replies reveal further fallacies in the Chinese Room argument, and reinforce the message of Turing's thesis.

The Robot Reply

This reply arises from the view that the person in the Chinese Room is prevented from understanding Chinese by a lack of sensorimotor connection with the reality that the Chinese characters represent. If the system were housed inside a robot that has sensory-motor capabilities, it would be able to "perceive" and "act," and thus, genuinely understand Chinese, since semantic value of information is heavily dependent on context.

Searle notes that the Robot Reply "tacitly concedes that cognition is not solely a matter of formal symbol manipulation, since this reply adds a set of causal relation with the outside world."[28] Like Searle's Chinese Room argument, the Robot Reply is an argument from semantics and emphasizes the role of understanding through perceiving and acting. The Robot Reply, therefore, cannot possibly defend Turing's thesis against Searle's objection, because it is vulnerable to an infinite recursion. As Searle explains, "Notice that the same thought experiment applies to the robot case. Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case..."[29] The Robot Reply is ineffective, but it remains popular.

Below, I will discuss why arguments from semantics fail to threaten Turing's thesis.

---

[28] Searle, "Minds, brains and programs."

[29] Ibid.

Reply to Searle's Argument from Semantics

Searle's Chinese Room argument focuses on semantic meaning as a necessary condition of cognition and understanding. He later formalizes this argument as follows:

(Axiom 1) Computer programs are formal (syntactic).

(Axiom 2) Human minds have mental contents (semantics).

(Axiom 3) Syntax by itself is neither constitutive of nor sufficient for semantics.

(Conclusion) Programs are neither constitutive of nor sufficient for minds.[30]

There are several ways in which one can unpack this argument to uncover its flaws. William Rapaport's approach, which he delineates wonderfully in his essay, *How to Pass a Turing Test*, is to claim that premise (S3) is wrong, and that syntax *is* sufficient for semantics. He explains that "what Searle alleges is missing from the Chinese Room is semantic links to the external world," for example, the link from some shape or squiggle referring to some meaningful thing, such as a hamburger (an example chosen by Rapaport).[31]

Rapaport points out that Searle makes two assumptions: that external links are necessary in order for the system to "attach" meaning to the appropriate symbols, and that computers have no means of linking to the external world; they only have access to what is internal to them.[32] First, if external links were necessary, it seems very well possible that computers could have them. This is the main point of the Robot Reply, as

[30] John R. Searle, "Is the Brain's Mind a Computer Program?," *Scientific American* 262 (1990): 27.

[31] Rapaport, "How to Pass a Turing test," 474.

[32] Ibid.

previously discussed. To review, the idea of the Robot Reply is that housing the Chinese Room inside of a robot "body" would give the system sensorimotor capabilities, which would allow the system to process information contextually rather than reflexively or mechanically. But, Rapaport asks, are these sensorimotor links necessary? He claims that they are not, and his argument is quite clever. He seeks to prove that insofar as the syntactic symbols are internal to a mind, the "semantic domain can be internalized."[33]

He recalls how we learn the meanings of words in the first place:

> How do I learn that 'tree' refers to that large brown-and-green thing I see before me? Someone points to it in my presence and says something like "This is called a 'tree'." Perhaps numerous repetitions of this, with different trees, are needed. I begin to associate two things, but what two things? A tree and the word 'tree'? No; to paraphrase Percy (1975: 43), the tree is not the tree out there, and the word 'tree' is not the sound in the air. Rather, my internal representation of the word becomes associated ("linked," or "bound") with my internal representation of the tree.[34]

When the reader imagines herself in the Chinese Room, it is likely that she sees a particular symbol more than once. She could, in fact, see a particular symbol many times, to the point where she remembers what it looks like when she comes across it and she remembers how to respond. Conceivably, she could remember numerous symbols in the same way, such that she would no longer need to look up how to respond. In a sense, she is actually internalizing the corresponding semantic domain for a given symbol, and learning Chinese. The syntactical symbols "should not be thought of as symbols *representing* something external to the system; although they *can* be related to other

---

[33] Ibid., 476.

[34] Ibid.

things by a third person, the only relations needed by the cognitive agent are all internal."[35]

I think that Rapaport is absolutely correct here, and his argument reminds me of the famous philosophical conundrum about the bent stick in the water. We believe that we genuinely see things, not mere illusions or mental images of them. This belief, however, becomes seriously problematic when we consider the bent stick in the water. When part of a stick is submerged in clear water, it undeniably appears bent, but when we remove the stick from the water completely, it becomes obvious that the stick is straight. If we believe that we see things as they really are, we must also believe that the fundamental nature of the stick changes sometime between its submerged-state and its removed-state. She must also concede that the stick itself has causal powers that result in our seeing it differently in water versus in air. In contrast, if one believes that all we see are mental images, the bent stick example makes perfect sense. When the stick is submerged in water, it doesn't appear bent because the stick is actually bent; it appears bent due to the way in which light refracts off of the surface of the water and onto the surface of the observers's eyes, where it is then processed as input by our fallible eyes and brains. Accepting this process eliminates the absurd notions that the stick has causal powers or that it fundamentally changes in certain circumstances.

The point is that context could very well be internalized, similar to our mental images of the submerged stick. It is important to note that my mentioning the eye in my description of the process by which a mental image is formed has nothing to do with the

[35] Ibid., 486.

eye as a means of perceiving the external world (as proposed in the Robot Reply). The eye could be replaced with anything that receives input, for example, a person in the Chinese Room.

Reply to Searle's Notion of Intentionality

To Turing, intentionality is completely irrelevant and in general, notions about theory of mind are too vague to serve as the foundation for any worthy argument. Those who wish to add anything valuable to the discussion about the nature of intelligence and thought should recognize that intentionality is a vague intuition, not a rigorous concept.

Searle is deeply anxious about the objective existence of consciousness, so he has to try to find ways to prove that the human mind is more than just a neurologically-realized machine. The "Chinese Room" is just a way of expressing an irrational belief that we are somehow exceptional, and that the universal principles of learning and intelligence we have worked out do not apply to us. Searle does not seem to understand that intelligence and consciousness are separable, and from a rigorous perspective, and function F that maps R $\rightarrow$ R can be considered "intelligent" provided it behaves according to some intelligible pattern (i.e. it passes the Turing Test).

These replies to Searle's Chinese Room argument are helpful to understand since some of them successfully defend and reinforce the validity of the Turing Test against Searle's claims. But it is important to note that Searle's argument does not affect the discussion of moral status of AI. When a rational actor confronts a machine that is able to pass the Turing Test, her moral stance towards that machine must be based on her

interaction with it, since we determine our moral stance towards other beings in this way. It's not whether or not an agent can feel pain, we must morally act as if it did, since we do not require an agent to prove that it is genuinely experiencing pain versus merely exhibiting behavior intended to "simulate" the experience of pain.

CHAPTER TWO

I. MORAL STATUS

In her notable bioethics article titled *On the Moral and Legal Status of Abortion*, Mary Anne Warren suggests that the following traits are "the most central to the concept of personhood,"

1. Consciousness (of objects and events external and/or internal to the being), and in particular the capacity to feel pain;
2. Reasoning (the developed capacity to solve new and relatively complex problems);
3. Self-motivated activity (activity which is relatively independent of either genetic or direct external control);
4. The capacity to communicate, by whatever means, messages of an indefinite variety of types, that is, not just with an indefinite number of possible contents, but on indefinitely many possible topics;
5. The presence of self-concepts, and self-awareness, either individual or racial, or both.[1]

[1] Mary Anne Warren, "On the Moral and Legal Status of Abortion," in *Biomedical Ethics*, 4th ed., ed. Thomas A. Mappes and David DeGrazia (New York: McGraw-Hill, 1996), 434-40, http://instruct.westvalley.edu/lafave/warren_article.html (accessed November 2, 2010).

Warren admits that formulating such precise definitions, as well as developing universal applications, often produces philosophically problematic results, so she tells us to consider these five criteria as a loose framework in which sufficiency and necessity are left open-ended. For example, in one's evaluation of whether or not some alien being is a person, one may find that traits (1) and (2) alone are sufficient to assess that being's personhood. However, Warren claims that a being that manifests *none* of the above traits is certainly *not* a person.[2] I believe that this framework for evaluating personhood is clear and intuitive, and sets an appropriate stage for discussion of the moral status of AI agents that can pass the Turing Test.

I do not think that it is contentious to say that one's passing the Turing Test provides evidence for one's manifestation of, or one's capacity to manifest, traits (3) and (4) at the very least. It seems fairly obvious to say that an AI that passes the Turing Test necessarily grasps a wide range of linguistic and conversational concepts at a level of sophistication that is comparable to that present in humans, that is, the capacity to communicate. Furthermore, we can assume trait (3) because of the isolated conditions in which the Turing Test is conducted. Neither the machine nor the machine's programmer has access to the interrogator's questions beforehand. It also seems intuitive to say that trait (2), likewise, can be assumed, due to the fact that it is required in order for a machine to be eligible for test participation in the first place. This refers to Turing's previously-mentioned specification of "digital computers" as the only eligible machine

[2] Ibid.

participants in the Turing Test, since they are modeled after and can closely mimic the functionality and complexity of "human computers."[3]

We are then left to ponder traits (1) and (5), and whether or not one's passing the Turing Test provides evidence for these traits, or if traits (1) and (5) are even suitable or appropriate conditions for evaluating the moral status of human beings, let alone the moral status of AI agents. Imagine an AI that makes the claim, "I am conscious." How can that machine prove itself? Would it be any easier for a human to prove herself if she were to make that claim? She might say, in her defense, that she has felt pain, she has felt joy, she has wondered, etc. But an AI could say these things just as easily, regardless of whether it is actually conscious or simply acting like it is. I believe that the only valid response to this is to bite the bullet; we don't require other humans to prove that they are actually conscious and not merely acting as if they are, so it seems unfair that we demand such proof from AI agents in order to award them moral status. Therefore, our social and moral assessments depend solely on appearance, namely, the appearance that one has intelligence, the appearance that one is conscious, the appearance that one has emotions, etc., rather than demanding reality. Mark Coeckelbergh, Professor of Philosophy of Technology at the Philosophy Department of the University of Twente, The Netherlands, explains that in human-to-human interaction:

> As a rule, we do not demand proof that the other person has mental states or that they are conscious; instead, we interpret the other's appearance and behaviour as an emotion. Moreover, we further interact with them as if they were doing the same with us. The other party to the interaction has

[3] Alan M. Turing, "Computing Machinery and Intelligence," Mind 59, No. 236 (1950): 436, doi:10.1093/mind/LIX.236.433 (accessed September 30, 2010).

> virtual subjectivity or quasi-subjectivity: we tend to interact with them as if our appearance and behaviour appeared in their consciousness.[4]

In accordance with this social norm of human-to-human interaction, it should follow that a sufficiently advanced AI -- that is, an AI that can pass the Turing Test, insofar as it can (in the very least) "imitate subjectivity and consciousness in a sufficiently convincing way" -- could also become a member of our moral community and "matter to us in virtue of [its appearance]."[5]

This phenomenological interpretation of morality (i.e. the appearance of consciousness, emotions, and mental states) not only helps us understand how we will likely interact with, and feel moral obligation towards, future AI agents, but it also helps us make sense of our current notion of human morality. I don't feel a sense of moral obligation towards my toaster, and I would never consider it a member of a moral community, simply because it doesn't appear to be conscious or experiencing joy. It's interesting to note that we hold similar attitudes towards some animals, but not towards others. On what basis are we awarding only certain animals with acceptance into our moral community? Certainly, if I were to beat my toaster with a hammer, or even shoot it with a shotgun, no one would protest my action on moral grounds. Likewise, if I were to harm an ant or a worm, I would also face no moral opposition. Yet why is our society morally opposed to canine abuse, for example? I believe that we ascribe moral agency to certain animals and not to others because some species give the appearance of having

---

[4] Mark Coeckelbergh, "Moral Appearances: Emotions, Robots, and Human Morality," *Ethics and Information Technology* 12 (2010): 238, http://www.springerlink.com/content/f6866544337v5822/fulltext.pdf (accessed November 10, 2010).

[5] Ibid.

mental states and emotions while others do not. As Coeckelbergh points out, "we treat a particular dog as a pet since it appears to have those emotions that make us see it as a companion."[6]

It is clear that the manifestation of consciousness and emotions is not required, and that the mere appearance of such is sufficient to draw a subject into our moral community, based on our notion of human morality. It is somewhat unsatisfying, however, to think that our concepts of moral obligation and responsibility are based purely on superficial principles. I will therefore attempt to reduce this property (one's ability to appear to be conscious or have emotions, or one's ability to appear to manifest Warren's traits (1) and (5)) to be merely a consequence of one's manifesting traits (2), (3), and (4), namely, one's intelligence.

Engineers and designers of AI agents, therefore, do not bear the burden of creating consciousness or emotions in their machines. If they aim to design a robot with traits (2), (3), and (4), it's quite plausible that the robot will learn to produce the appearance of having consciousness, emotions, and self-awareness in relation to its environment and other entities occupying that environment.

II. BRAITENBERG'S VEHICLES

[6] Ibid., 239-40.

This idea of evolutive behavior in machines was most famously illustrated by Valentino Braitenberg in his revolutionary book, Vehicles: Experiments in Synthetic Psychology. The vehicles he discusses are extremely simplistic in their programming and design. Each vehicle just has a set of sensors connected to its motor, similar to a living creature's neurological connections to its eyes or ears. With these sensors, the vehicle can interact with its environment and exhibit increasingly complex behaviors as a result of its sensor-actuator connections. Eventually, the vehicle appears to be capable of expressing fear, aggression, love, etc.[7]

Braitenberg's vehicles are rudimentary implementations of AI. They have intelligence but lack cognition, yet they can adapt to shifting environmental conditions. Their behavior, which results from mere light detection, is undeniably goal-oriented. The vehicles appear to have the same level of intelligence as cockroaches, but they do not undergo any cognitive processes, and they are instantiated programmatically.

The fact that intelligent-type behavior develops naturally (i.e. without being explicitly programmed) in Braitenberg's vehicles is not only intriguing, but also promising in terms of feasibility of advancements in AI.

---

[7] Valentino Braitenberg, *Vehicles: Experiments in Synthetic Psychology* (Cambridge: MIT Press, 1984).

CONCLUSION

The ability to communicate effectively with humans is by far the biggest obstacle for AI engineers and designers who seek to build non-human citizens. Linguistic nuances, such as slang terms and idioms, typically confuse AI systems. On the other hand, current AI systems compute and strategize faster and more accurately than we ever could. Future projects in this field will undeniably have a huge impact on our lives. Perhaps robots will fight our wars, take our jobs, save our lives, etc. They might even become the Hummingbird Hawk-moths to the human species.[1]

As rational actors, we must base our moral stance towards AI agents solely on our interactions with them. Therefore, machines that can pass the Turing Test will likely join our moral community to some degree. We will have no reason to treat them differently than other beings that exhibit human-type intelligence. Questions regarding the moral responsibility of these agents will be no more answerable or unanswerable than questions regarding the moral responsibility of humans.

[1] The Hummingbird Hawk-moth serves the same ecological role as the hummingbird, and is said to be a byproduct of convergent evolution.

BIBLIOGRAPHY

Braitenberg, Valentino. *Vehicles: Experiments in Synthetic Psychology*. Cambridge: MIT Press, 1984.

Chang, Kenneth. "NASA's Quest to Send a Robot to the Moon." *The New York Times*, November 2, 2010, Science section, New York edit ion. http://www.nytimes.com/2010/11/02/science/space/02robot.html (accessed November 2, 2010).

Coeckelbergh, Mark. "Moral Appearances: Emotions, Robots, and Human Morality." *Ethics and Information Technology* 12 (2010): 235-41. http://www.springerlink.com/content/f6866544337v5822/fulltext.pdf (accessed November 10, 2010).

Copeland, Jack B. "The chinese room from a logical point of view." In *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John M. Preston and Michael A. Bishop, 109-122. Oxford: Oxford University Press, 2003.

McElroy, Molly. "I Want to See What You See: Babies Treat 'social Robots' as Sentient Beings." *University of Washington News*, October 14, 2010. http://uwnews.org/article.asp?articleID=60848 (accessed November 2, 2010).

Mone, Gregory. "The New Face of Autism Therapy." *Popular Science*, June 1, 2010. http://www.popsci.com/science/article/2010-05/humanoid-robots-are-new-therapists (accessed November 5, 2010).

Palmer, Jason. "Call for Debate on Killer Robots." *BBC News*, August 3, 2009. http://news.bbc.co.uk/2/hi/technology/8182003.stm (accessed November 5, 2010).

Rapaport, William J. "How to Pass a Turing Test." *Journal of Logic, Language and Information* 9 (2000): 467-490.

Searle, John R. "Is the Brain's Mind a Computer Program?" *Scientific American* 262 (1990): 26–31.

Searle, John R. "Minds, brains and programs." *Behavioral and Brain Sciences* 3 (1980): 417-57. doi:10.1017/S0140525X00005756 (accessed September 30, 2010).

Thrun, Sebastian. "What We're Driving at." *The Official Google Blog*, October 9, 2010. http://googleblog.blogspot.com/2010/10/what-were-driving-at.html (accessed November 2, 2010).

Turing, Alan M. "Computing Machinery and Intelligence." *Mind* 59.236 (1950): 433-60. http://mind.oxfordjournals.org/content/LIX/236/433.full.pdf+html (accessed September 30, 2010).

Warren, Mary A. "On the Moral and Legal Status of Abortion." *Biomedical Ethics*. Ed. Thomas A. Mappes and David DeGrazia. 4th ed. New York: McGraw-Hill, 1996. 434-40.

Weizenbaum, Joseph. "ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM* 9.1 (1966): 36. http://web.archive.org/web/20071026055950/http://www.fas.harvard.edu/~lib51/files/classics-eliza1966.html (accessed November 24, 2010).

Wiener, Norbert. *Cybernetics or Control and Communication in the Animal and the Machine*. 2nd ed. Cambridge: MIT, 1965.

Woollacott, Emma. "First All-robotic Surgery and Anesthesia Performed." *TG Daily*, October 20, 2010. http://www.tgdaily.com/general-sciences-features/52099-first-all-robotic-surgery-and-anesthesia-performed (accessed November 2, 2010).

FURTHER READING

Block, Ned. "Searle's arguments against cognitive science." In *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John M. Preston & Michael A. Bishop. Oxford: Oxford University Press, 2003.

Churchland, Paul M. and Patricia S. Churchland. "Could a Machine Think?" *Scientific American* 262 (1990): 32-37.

Dennett, Daniel Clement. *Kinds of Minds*: Toward an Understanding of Consciousness. New York: Basic Books, 1996.

French, Robert M. "Subcognition and the Limits of the Turing Test." *Mind* 99 (1990): 53-66.

Hauser, Larry. "Searle's Chinese Box: Debunking the Chinese Room Argument." Minds and Machines 7 (1997): 199-226.