



Philosophy of Consciousness And Ethics in Artificial Intelligence

By Shawn Kilmer
December, 2007

Philosophy of Consciousness And Ethics In Artificial Intelligence

By Shawn Kilmer
University Of Oregon
December 2007

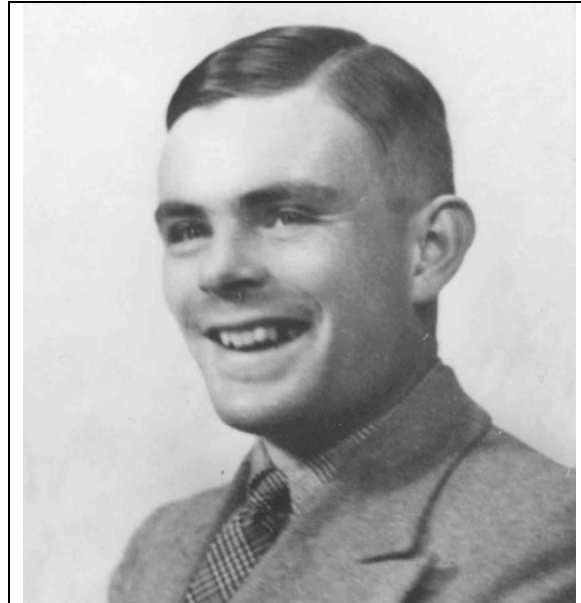
Front Cover: Which of the two beings depicted on the front cover should be granted more rights than the other? What of the CPU circuit depicted? If it is capable of simulating all the same behaviors of an animal or even a human, should it also be granted rights? The picture of the man is actually a computer generated 3D model from the NVIDIA Corporation, while the cow is indeed a photo of a real cow. This is just to show how we can be fooled in what gives us the impression of possessing intelligence.

Since the creation of the first 'smart' computation machines in the time of Alan Turing, the father of computation and artificial intelligence, questions have risen over what makes a machine intelligent. Debates have sprung up over the decades on many questions regarding the differences in the consciousness of humans and the possible consciousness of machines – both in current machines and yet-to-be-invented ones. Some of the questions raised are: what is the minimum definition of what we would consider consciousness in any entity, machine or biological? If we do at some time consider a machine intelligent or as having a conscience, are there any ethical implications raised with how these machines are used? If machines can meet the Turing Test or meet a definition of consciousness that we could come up with, should they be granted some of the rights that other living entities have? Will we need to modify our laws about human and animal rights if machines can meet these requirements?

MINIMUM DEFINITION OF CONSCIOUSNESS

In starting to think about and address the question of the minimum definition of what we would consider consciousness, we

first need to come to some understanding about what we mean



Alan Turing ²

by the term consciousness. Let's first get the opinion of Turing himself: "I do not wish to give the impression that I think there is no mystery about consciousness... But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper" (Alan Turing, quoted from Russel, 953). I quite agree with this stance on the present question, and that is that no one currently asserts that they have a perfect understanding of consciousness. Given that, we need to decide what consciousness will mean for us in a given application.

I take my definition of consciousness herein to follow along the lines of a quite common response to the famous Chinese Room Problem proposed by John Searle. In summary,

Searle argues against “Strong AI,” which is the theory that machines can have a conscience and be intelligent. To illustrate his point, he describes a room in which a non-Chinese-speaking man receives input through a slot in the form of written Chinese characters. He has a rule book written in the language he understands, English, that tells him when given certain Chinese characters which Chinese characters he should arrange on another sheet to pass back out through the slot in response. By following the rule book, the man would be able answer questions correctly in Chinese without having the faintest understanding of the language (Searle, 1980).



Chinese Characters ³

Numerous responses from many different disciplines have been written in response to the given paradigm, and I fall into the school of thought of the simulator

response: If every question that we could pass into the room produces a correct answer back, then we would have a perfect simulation of someone in the room who did in fact understand Chinese. The operator within the room still may have no understanding of the language, but as a unit combined with the rule book and the room and the characters, the *room* does understand Chinese. Since we cannot distinguish between understanding and perfectly simulated understanding, they are non-different.

I apply this response to the question at hand of consciousness. If we have managed to create a machine that can perfectly simulate what we understand as consciousness, then it is non-different from actual consciousness. I fully agree with Terry Winograd’s ideas of whether or not we will have created consciousness if we manage to construct exactly an artificial brain, piece for piece:

“Would it think? Would it be conscious? Well, I’m ultimately a materialist, so I would say of course. If you really could duplicate it piece by piece, it would be all the same pieces; there’s no ethereal soul that makes me have consciousness--it is in the physical properties of my brain and my nervous system. So if you take that very broad notion of computer, then it becomes a matter of whether you’re a materialist or not” –Terry Winograd (Koch).

Ray Kurzweil weighs in on the question by adding a remarkable mental exercise of his own that goes like this: Suppose you know a guy named Jack. Jack gets a prosthetic medical replacement to one of his body parts after an injury. Is he still the same Jack to you that you've known? Of course he is. Now suppose one by one, he began replacing body parts with prosthesis (assuming that the medical technology existed to replace every possible body part). He replaces his arms, even eyes, and auditory system, until he has even replaced his brain neuron by neuron to be an exact prosthetic analogue, with each piece functioning exactly the same as the original part did. He has the same sense of humor and same stories to recall with this reconstructed model, so is it the same Jack (Kurzweil, 53)?

Kurzweil's answer seems to be a resounding 'yes,' as he concludes by citing the 'consciousness-is-just-a-machine-reflecting-upon-itself' school of thought. This school of thought defiantly throws aside the concepts of consciousness and free-will just as I so longingly wished I could have done during a year's worth of college philosophy courses: "Consciousness and free will are just illusions induced by the ambiguities of language... We can build that in a machine:

just build a procedure that has a model of itself, and that examines and responds to its own methods. Allow the process to reflect upon itself... Now you have consciousness" (Kurzweil, 58).

Jeff Hawkins points out that "most people think consciousness is some kind of magical sauce that is added on top of the physical brain... In this view, consciousness is a mysterious entity separate from brains" (Hawkins, 159). He says "many people persist in believing that consciousness is different and can't be explained in reductionist biological terms...I believe that consciousness is simply what it feels like to have a neocortex"(Hawkins, 160), and continues to provide his theory for why people believe consciousness to be an unexplainable separate entity:

"The cortex has no ability to model the brain itself because there are no senses in the brain. Thus we can see why our thoughts appear independent of our bodies, why it feels like we have an independent mind or soul. The cortex builds a model of your body [and the world around you] but it can't build a model of the brain itself. Your thoughts, which are located in the brain, are physically separate from the body and the rest of the world. Mind IS independent of the body, but not of the brain" (Hawkins, 200).

ETHICAL USAGE FOR INTELLIGENT MACHINES

If we do at some time consider a machine intelligent or as having a conscience, are there any ethical implications raised with how these machines are used? Again, in addressing this question, *ethics* is a term that really needs to be examined. When asked about intelligent machines having moral sense or following ethical rules, prominent artificial intelligence scientist and author of The Emotion Machine Marvin Minsky responded by saying he wasn't sure there was any such thing called moral sense or ethical rules (*Philosophy Talk*, May 20 2007). He said in his view, people just define ethical rules by expressions of 10 or so if-then statements in their minds about a given issue, so there is no reason a machine cannot express the same arbitrary rules.

I am under the same impression, and I do not think it is currently a popular one. I believe it is not a popular opinion to have on ethics for the same reason that things like stem cell research and even evolution have so many people speaking against them in the present days: because of religious dogmatism that so often is blind to and impedes scientific theory. Christof Koch understood this when he says "There's a

sense that once you start talking about consciousness, next you'll talk about religion."¹

For the sake of trying to avoid that discussion, I will suffice it to say that whether or not there are real things called ethics and morals that actually mean anything, people can only talk about these concepts by what they observe of them rather than what they know to be factually true about their nature. Since we have different personal stances on what ethical and moral behavior are all across the world (in fact, I would imagine we could find every possible stance), we know that they are not same thing to all people. It follows that therefore our conception of *ethics* amounts to not much more than what a given group of people decide is ethical behavior within their group –which, as Minsky put it, is just a series of if-then definitions.



Marvin Minsky ⁴

Since as of yet we are still having problems producing computers that can entirely pass the Turing Test put forth over 50 years ago,

we still are not at a point where we really have to worry about the ethical considerations of human-conscience-like machines. Our computers are not calling out to us in genuine desperation to not shut them off, nor to be our friends or lovers. I don't think anyone feels guilty when we shut down our desktop computer at night or even when we slaughter innocent simulated civilians while playing Grand Theft Auto™. This is because of a very important fundamental difference in machine consciousness versus living consciousness which I will point out in response to the next question.

MACHINE RIGHTS

If machines can meet the Turing Test or meet a definition of consciousness that we could come up with, should they be granted some of the rights that other living entities have? Firstly, there is a gaping discrepancy in trying to make this comparison. Machines and machine-consciousness currently have the advantage of having the ability to have their entire state stored before turning them off, allowing for complete and non-destructive restoration when power is reapplied. Humans, however, have not yet figured out the technology to make perfect backups of our brain state, nor

to restore dead bodies into which to load the saved brain states.

Would it be ethical to 'pull the plug' on these artificial beings which simulate consciousness? This gets back to Kurzweil's analogy of Jack. He extends the analogy to the scenario of being able to do a complete backup of Jack's entire brain state neuron for neuron, and being able to save it digitally, but having this being possible only by means of a destructive process. The biological matter that composed Jack would be entirely destroyed, but we would have his entire new prosthetic body and the brain state backup ready to load into it to produce a perfect simulation of the previous Jack.

Kurzweil then says if a person thinks this destructive process would be unethical, that "therefore, it can be argued that the Star Trek characters are committing suicide each time they teleport, with the new characters being created. These new characters, while essentially identical, are made up of entirely different particles... Is consciousness a function of the actual particles or just of their pattern and organization?" (Kurzweil, 53). While it is hard to comprehend our reactions to this situation since we do not actually have a way of 'backing up' a human brain-state, if we did I think we would have to come to see why this would not be an unethical practice – for the same reason that turning off your computer is not.

If the only thing that ethically sets apart intelligent machines and life is the ability to have backups created and be restored, and humans could become backed up and restored with substitute components (though requiring the destruction of the biological components) then it follows that it would not then be unethical to destroy the biological components. Though this seems to be a logically sound position, I am certain that it wouldn't go over well with those who hold religious beliefs since they tend to disregard logic.

IMPLICATIONS FROM EXISTING LAWS

Will we need to modify our laws about human and animal rights if machines can meet all the same requirements? The question of animal rights is an interesting one here, because clearly machines are regarded by some much more highly and personally than some regard animals. I have discussed the fundamental and important difference between living and machine consciousness and from the conclusions that follow it seems that all living consciousness should be of greater 'ethical' concern to us than any machine consciousness because of the condition we are under of not having the technology yet to back up brain-states.

It is an important point which I think

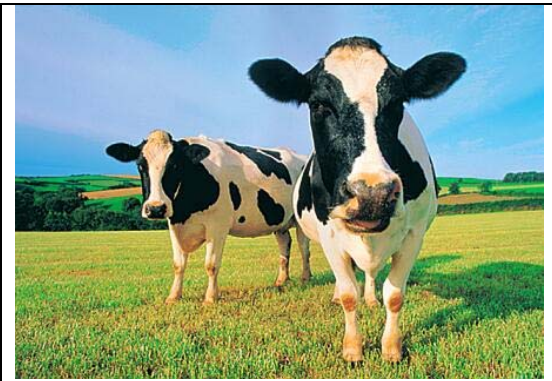
deserves careful consideration by everyone. Renowned scientist Roger Penrose discusses the point that physiologists have argued that the reticular formation structure of the brain might be taken to be the seat of consciousness, if such a seat exists. He points out it is a very ancient part of the brain evolutionarily, and if that is all that is necessary for consciousness, "then frogs, lizards, and codfish are conscious"(Penrose, 494). He continues,

"What evidence do we have that lizards and codfish do NOT possess some low-level form of consciousness? What right do we have to claim, as some might, that human beings are the only inhabitants of our planet blessed with an actual ability to be 'aware'? Are we alone, among the creatures of the earth, as things whom it is possible to 'be'? I doubt it."

"...Many philosophers and psychologists seem to take the view that human consciousness is very much bound up with human language... It is language, according to this view, that distinguishes us from other animals, and so provides us with our excuse for depriving them of their freedom and slaughtering them when we feel that such need arises"(Penrose, 495)

For Jeff Hawkins, AI scientist and creator of the PalmPilot and Treo SmartPhone, there is no dispute that all living things are intelligent by virtue of the two standards: memory, and prediction about the structure of the world for

their benefit of reproduction. All life possesses these two intelligent traits, they just range across a whole “continuum of methods and sophistication in how they do it” (Hawkins, 179). He even includes plant life in this continuum- they ‘remember’ what worked in predicting the structure of their world for benefit of reproduction by way of DNA that gets passed on with the most successful traits. This is a non-conscious ‘memory’ or prediction which they cannot control over their own lifetimes, but it is one nonetheless and earns them their place at the low end of his continuum. In this way, “All mammals, from rats to cats to humans... are intelligent, but to differing degrees” (Hawkins, 180).



Cows ⁵

I think we need to reflect upon the points herein, and before these ‘ethical’ considerations about machines become real and in current practice, we should recognize

that the fundamental difference still exists between all life and machines –our power switches cannot be turned back on... yet. It is a bit confusing and disconcerting that so much debate constantly takes place over the ethical implications of machines in the scientific community as well as in the general media, when very little widespread attention is given to the atrocious amounts of animals that are being exploited for our benefit in ways in which compete with Auschwitz in their level of brutality and disgust. All of this talk on the issue while the giant difference remains –machines can be restored and backed-up while animals such as codfish, lizards, cows, and human beings cannot. I believe it is also, once again, largely due to religious belief and held tradition that keep this issue from getting the immediacy of attention it deserves. People tend to not want to change or call into question standing practices.

CONCLUSIONS

When we arrive at the technological epoch of machine-consciousness that is a perfect and indistinguishable simulation of human consciousness, then yes, we may need clauses in our laws defining rights to living beings that distinguish between entities which can be non-destructively restored and those which cannot. When we are capable of creating such consciousness, the technology would perhaps

soon follow that would allow the gap to be closed on animals' limitation of not being able to have saved states of themselves, and actually allow backups – destructive or not. Then, questions of ethics become entirely new, needing to be rethought, with some of our old “if-then statements” becoming obsolete. If we have a perfect simulation of Jack's consciousness, and perfectly simulated artificial parts for Jack, then for my money he would be the same old Jack to me.

In the mean time, it seems that the entire emphasis of debate needs to be shifted. Away from the moral and ethical implications of increasing technology and 'artificial' intelligence in our society, and to the ethical and moral implications of our current actions in our lives and the world, starting with an honest, scientific re-evaluation of what ethics and morality are in themselves, for all life. ◀

FOOTNOTES

¹ Nearly my whole interest in writing this paper was summarized by this passage I found in Russel and Norvig, regarding why people may have ethical concerns over the advance of AI:

"people might lose their sense of being unique. Weizenbaum (1976) ... points out some of the potential threats that AI poses to society. One of Weizenbaum's principal arguments is that AI research makes possible the idea that humans are automata -- an idea that results in a loss of autonomy or even of humanity. We note that the idea has been around much longer than AI, going back at least to L'Homme Machine (La MMetrie, 1748). We also note that humanity has survived other setbacks to our sense of uniqueness: De Revolutionibus Orbium Coelestium (Copernicus, 1543) moved the Earth Away from the center of the solar System and Descent of Man (Darwin, 1871) put Homo Sapiens at the same level as other species. AI, if widely successful, may be at least as threatening to the moral assumptions of 21st centure society as Darwin's theory of evolution was to those of the 19th century."(961)

² Source: http://en.wikipedia.org/wiki/Image:Alan_Turing.jpg

³ Source: <http://images.jupiterimages.com/common/detail/18/59/23325918.jpg>

⁴ Source: <http://philosophytalk.org/pastShows/ArtificialIntelligence.html>

⁵ Source: <http://www.poster.net/anonymous/anonymous-cows-5000235.jpg>

Works Cited

- Hawkins, Jeff. On Intelligence.
New York: Times Books, 2004
- Henig, Robin Marantz. "The Real Transformers." The New York Times. 29 July, 2007.
Accessed November 2007
(<http://www.nytimes.com/2007/07/29/magazine/29robotst.html>)
- Interview with Marvin Minsky. Interviewers: Ken Taylor and John Perry. Philosophy Talk
(<http://philosophytalk.org/pastShows/ArtificialIntelligence.html>). 20 May, 2007
- Koch, Christof. "What Is Consciousness?." Discover. 1 November, 1992.
Accessed November 2007
(<http://discovermagazine.com/1992/nov/whatisconsciousn149/>)
- Kurzweil, Ray. The Age Of Spiritual Machines.
New York: Penguin Group, 1999
- Penrose, Roger. The Emperor's New Mind.
New York: Oxford University Press, 1989
- Russel, Stuart, and Peter Norvig. Artificial intelligence: A Modern Approach, Second Edition. New Jersey: Prentice Hall, 2003
- Searle, John. "Minds, Brains and Programs", Behavioral and Brain Sciences 3 (3): 417-457. 1980