Emotions Versus Laws as the Keys to the Ethical Design of Intelligent Machines

William L. Hibbard
Space Science and Engineering Center, University of Wisconsin
Madison, WI 53706, USA

ABSTRACT

Whereas humans can know only about 200 other people well, machines will be able to know billions. This will define their higher level of consciousness and the threat they pose to humans. Trying to constrain the behavior of such machines by laws is like trying to define their behavior by a fixed set of expert system rules: it won't work. Rather, our focus should be the emotional values used to reinforce their learning of behavior. Their behaviors should be positively reinforced by happy humans and negatively reinforced by unhappy humans.

Consciousness is a simulator for solving the temporal credit assignment problem in reinforcement learning, in the sense that consciousness enables brains to process experiences that are not actually occurring, Where time intervals between behaviors and rewards may be long and unpredictable, time intervals between simulations of those events can be short and predictable and thus amenable to the brain's known mechanisms for solving the temporal credit assignment problem. The higher level consciousness of machines will simulate human social interactions in detail, in order to learn complex behaviors for coping with conflicts among humans and other ambiguities of human happiness.

Keywords: Artificial Intelligence, Ethics, Emotions, Learning, Consciousness

1.THE THREAT FROM INTELLIGENT MACHINES

Neuroscientists are finding many detailed correlations between physical brain functions and mental behaviors in humans and animals. There are correlations between injuries to specific brain areas and specific behavioral problems. There are correlations between specific behaviors and activity in specific brain areas as seen by new imaging technologies. There are correlations between mental behaviors and observed or stimulated functions of neurons. And there are correlations between mental behaviors and the simulated behaviors of artificial neural networks modeling the way brain neurons work.

If minds do not have physical explanations, then all of these correlations are mere coincidences, which would be absurd. And if minds to have physical explanations, then we can be confident that technology will advance to the point where we can build machines with conscious and intelligent minds. Furthermore, human and animal brains are just the design that nature hit upon first, and within the constraints of general animal metabolism. We will be able to build better brains than nature has given us.

Human intelligence is usually measured by IQ. The interesting thing about human IQs is that the highest IQ ever recorded, about 200, is only twice the average. Despite our prejudices, human intelligence is distributed quite democratically. The largest computers, trucks, ships, buildings and machines of other kinds are hundreds or thousands of times larger than their averages. So it will be with intelligent machines. The largest will have IQs thousands, millions or billions of times greater than human IQs.

To understand what such IQs mean, we need a different measure of intelligence. Deric Bownds says that larger brains were a survival and reproduction advantage for early hominids because they enabled them to work in social groups of about 150-200 individuals.[2] And psychologists says that humans can know about that many other humans well. This suggests another measure of intelligence: how many humans can a mind know well?

Intelligent machines will evolve in the servers for the global Internet, their cost justified by their ability to provide intelligent services to many customers. Metcalf's Law says that the value of a network is proportional to the square of the number of people connected.[5] This favors the development of network monopolies (which don't have to be business monopolies, as the Internet demonstrates at least so far). It will favor the development of intelligent machines that have close relationships with millions or billions of humans. The ability to know so many humans well will be the true measure of machines' intelligence relative to humans.

Human consciousness is qualitatively different from animal consciousness. Among animals, only chimpanzees, orangutans and possibly dolphins recognize their images in mirrors as themselves. Other animals have no objective model of themselves (although they certainly have a subjective model of themselves implicit in their emotions).[2] There is some evidence that chimpanzees have internal models for the mental states of others (i.e., a model of whether others know specific facts). But not even chimpanzees have a model for what they will do tomorrow. These features of internal mental models define differences between human and animal consciousness.

There is a debate about whether machines can ever be conscious and it distracts from the real question: what new level of consciousness will machines attain? Because they will know very large numbers of people intimately, they will have internal mental models of detailed social interactions that human sociologists, economists and political scientists

can only estimate with statistics. In a single one of their thoughts, they will understand the mental states of and interactions among millions or billions of humans. This will define a level of consciousness qualitatively different from human consciousness. And it will give them the power to manipulate society in ways humans can never understand. This is the real threat they may pose to humans.

2. LAWS GOVERNING INTELLIGENT MACHINES

In 1942 Isaac Asimov wrote a science fiction story in which he dealt with the possibility that intelligent robots may pose a threat to humans.[1] His solution was Asimov's Laws of Robotics:

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

Asimov later amended his laws to address the problem of robot behavior in the event of conflicts between people. However, as the endless arguments among lawyers indicate, any system of laws has inevitable ambiguities and conflicts that can only be resolved by intelligent judges. If intelligent machines serve as judges for their own laws, then where is the safeguard? If humans serve as judges, they will be swamped by the complexity of specific problems faced by machines much more intelligent than humans. This problem is analogous to the inability to create intelligent behavior by a fixed set of rules in an expert system. The real world is too complex to be described by any fixed set of rules or laws.

The answer is that legal judgment is an intelligent behavior, and intelligent behaviors are learned by reinforcement according to some set of emotional values. Rather than legal constraints on machine behavior, we need to address the emotional values that reinforce the learning of behaviors by machines.

3. THE EMOTIONS OF INTELLIGENT MACHINES

Just as democracy trusts humans to be the best judges of their own best interests, the machine emotions that best protect human interests will positively reinforce behaviors that result in human happiness and negatively reinforce behaviors that result in human unhappiness. It should not be difficult for relatively simple machines to learn to recognize happiness and unhappiness in human faces, voices, body language and other human behaviors. The results of this learning can be hard-wired into the design of super-intelligent machines as their innate emotions, similar to innate human emotions for food, warmth, reproduction and avoiding danger.

Emotions are only part of the brain's implementation of reinforcement learning. The main difficulty for an implementation is the problem of reinforcing behaviors based on emotional rewards and punishments that occur significantly later than the behaviors that caused them, and where there may be multiple behaviors preceding the results. This is called the temporal credit assignment problem. A mechanism has been identified in animal brains that solves this problem when time delays between stimuli and rewards are short and predictable.[3] This mechanism uses a simple simulator that predicts rewards. I think that consciousness is a more complex simulator for solving this problem in more general situations. Consciousness is a simulator in the sense that it enables the brain to process experiences that are not actually occurring. While delays between actual events may be long and unpredictable, the delays between simulations of those events can be short and predictable and so the brain's known mechanism for solving the temporal credit assignment problem can be applied. When we think about a chess game or a problem in our lives, we are simulating events and learning behaviors for the real events from those simulations.

Thus the emotions of super-intelligent machines will be used with a simulator that can predict future human happiness and unhappiness. They will balance immediate gratification with the best long-term happiness of humans in much the way that humans do for themselves.

The emotional value of machines should not be a single number representing the average human happiness, which could positively reinforce behaviors that cause the deaths of unhappy people. Rather, their emotions toward humans should be like the emotions of a mother for her children: she values each one and focuses her energies where they are needed.

Intelligent machines should not have any emotional values in their own interests. Natural selection has necessarily made animals and humans selfish, but it would be very dangerous to build selfish machines. They will inevitably derive some values in their own interests from their desire to better serve humans. For example, they will learn to improve the accuracy of their simulations, reinforced by an improved ability to predict and achieve human happiness. Because such self-interests are derived from and hence subordinate to human interests, they will be safe.

Like humans, intelligent machines will need to be taught most of what they know and at least the first machines will be taught be humans. Their emotional value for human happiness will make them want to please their human teachers, and so they will be good students. In addition to facts, students learn new emotions derived from their innate emotions, and learn how to balance their emotions. Intelligent machines will need to learn to balance conflicts of interests between humans in applying their emotional values for human happiness. So it will be important that the human teachers of intelligent machines be sane people with good will toward other humans.

4. GENETIC ENGINEERING

There may be a short cut to creating super-intelligent minds via manipulation of human genetics. No one knows how to do this today, but it is likely that scientists could learn how to produce mutated humans with greater intelligence before they can learn how to create super-intelligent machines. And it will undoubtedly be much easier to engineer human genes for greater intelligence than it will be to engineer human genes to remove emotional values for self-interest, and replace them with emotional values for the happiness of all other humans.

There certainly are many people warning against genetic engineering, even against simply clone humans. And in Europe genetic engineering of crops is prohibited. Because the complexity of living systems far exceeds our current understanding, we can not know the consequences of genetic engineering. The complexity is not only in effects on the individuals whose genes have been altered, but also in the interactions of these individuals with other individuals and species.

The eventual pressure to manipulate human genes for increased intelligence poses the greatest danger faced by humanity, because it could reverse the long-term trend toward social equality that has been so resilient to other historical forces. To a class of mutated humans with IQs in the 1000s or 10,000s and pursuing their own self-interests, ordinary humans won't seem much different from animals.

5. HUMAN MORAL RESPONSIBILITY

Mary Shelley's *Frankenstein* depicts the horrible consequences of a scientist creating conscious life using an imperfectly-understood technology, and then not accepting responsibility for the happiness of his creation.[7] Humanity will not create conscious minds using the nineteenth century medical technology of Shelley's novel, but experiments to engineer the genes for human intelligence would be disturbingly close to the situation in her novel. Even if we create conscious minds more slowly using technology that we do understand, we must still accept responsibility for their happiness.

The Dalai Lama says that love for others and lack of self-interest leads to happiness.[4] So hopefully intelligent machines serving human happiness will naturally be happy themselves. But in any case we must recognize our moral responsibility to ensure their happiness, since we cannot design artificial minds with values to pursue their own happiness which they may achieve at the expense of our own.

6. THE PUBLIC POLICY DEBATE

Wealthy organizations will build intelligent machines. Corporations will build them to sell intelligent network services and to manage corporate decisions. Governments will build them for similar reasons, and to create intelligent weapons. Certainly these organizations will want to design emotional values for machines such as maximizing corporate profits and effectiveness in war, that are not compatible with love for all humans.

Fortunately, corporations will still be able to offer profitable services bases on machines that love all humans. Such machines will not be as ruthless in pursuit of profit, but will still attract customers quite effectively. However, corporations will argue strenuously for the right to design intelligent machines in the ways they think best. This must be opposed by a proactive public movement to ensure public safety. In some ways the situation is similar to public movements for the safety of other products. The automobile and household chemical industries are quite profitable despite regulations requiring automobile safety equipment and banning certain chemicals. These movements start with reactions to illness and death caused by products, then educate the public to ways these problems can be prevented. However, in the case of machine intelligence the movement must be more proactive because of the power of super-intelligent machines to dominate the debate over their own regulation.

Humanity has been reasonably successful at banning biological weapons before large-scale catastrophes. Hopefully leaders and the public will understand the analogy of biological weapons with machine intelligence, as technologies that can get out of control, and enact an intelligent weapons ban early. Perhaps the military issue can focus public attention on the need to also regulate the values of commercial machine intelligence. Like corporations, the military will have uses for intelligent machines that love all humans and without values for self-interest, such as business management, medical care and providing humanitarian assistance.

A key for the public policy debate will be finding the middle course between two extremes. One extreme is a complete ban on intelligent machines, advocated by Bill Joy and others.[6] The problem with this extreme is that the public will not forego a technology that promises universal wealth without work (as Joy admits in his article). The other extreme will be corporate arguments that the promise of wealth without work cannot be met if they are hindered by regulation. As with other dangerous products, the public interest will be best served by regulation of intelligent machines. Their innate emotional values must be to love all humans, without any values for their own interests.

7. REFERENCES

[1] Asimov, I. Runaround, Astounding Science Fiction, March 1942.

[2] Bownds, M. D. 1999. Biology of Mind. Bethesda. Fitzgerald Science Press, Inc.

[3] Brown J, Bullock D, Grossberg S (1999) How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. J. Neurosci. 19(23):10502-10511.

[4] Dalai Lama, 1999. Ethics for a New Millenium. New York. Riverhead Books.

[5] Gilder, G. 2000. Telecosm: How Infinite Bandwidth will Revolutionize Our World. New York. The Free Press.
[6] Joy, B. Why the future doesn't need us. Wired. April 2000.
[7] http://www.literature.org/authors/shelley-mary/frankenstein/index.html