Representations of the Brain and the Mind in Computer Science: Is there Room for Religious Concepts?

Lundy Lewis

Department of Information Technology, Southern New Hampshire University
2500 North River Road
Manchester, NH 03106-1045
l.lewis@snhu.edu

Abstract

In this paper we review the essential concepts of representing brain and mind behaviors as computer programs. First, we review the essence of neural networks, rule-based reasoning systems, and case-based reasoning systems and summarize the progress of these representations to date. Next, and more importantly, we ask a series of interrelated questions from a religious point of view: Is there room for religious concepts in these representations? If so, are representations of religious concepts in computational brain and mind theories consistent with our ordinary understanding of religious beliefs and attitudes? Does the representation of religious concepts in computer programs afford insights into the study of computer science, and conversely, does such representation afford insights into the study of religion?

Introduction

Computer Science has made important strides in recent years towards representing the brain and the mind as computer programs. For example, artificial neural networks are approximations of brain activity, while rule-based reasoning systems are approximations of mind activity. An important assumption of this work is that the brain and the mind are distinct, but interrelated entities, which raises the problem of characterizing the precise interaction between them. A popular theory of such interaction is epiphenomenalism – the view that mental activity is a side-effect, or a function of, brain activity.

In this paper we review the essential concepts of representing brain and mind behaviors as computer programs. First, we review the essence of neural networks, rule-based reasoning systems, and case-based reasoning systems and summarize the progress of these representations to date. Next, and more importantly, we ask a series of interrelated questions from a religious point of view: Is there room for religious concepts in these representations? If so, are representations of religious concepts in computational brain and mind theories consistent with our ordinary understanding of religious beliefs and attitudes? Does the representation of religious concepts in computer programs afford insights into the study of computer science, and conversely, does such representation afford insights into the study of religion?

Interesting possibilities and questions seem to be embedded in this topic: First, as we model the intelligence of our advanced robots after humans, should we expect them to exhibit religious behaviors? That is, might there be a level of computational intelligence at which we could have reasons to expect, say, religious questions to emerge naturally? On the side of brain research, after all, it has been argued that religious ideas and experiences arise unavoidably from human brain activity. Accordingly, if we model robotic intelligence well enough on human brain activity, perhaps advanced robots could be expected to form religious ideas of their own. An alternative line of thought challenges the foregoing suggestions and argues that no religious features would likely arise spontaneously from computational intelligence. Along these lines, one might question what reason we might have to incorporate religious representations in the brain or mind of a robot, and ought we incorporate such representations even if we did have a good reason?

These questions may invoke deeper philosophical questions regarding the current state of computational brain and mind studies. For example, we may return to the assumption that brain and mind behaviors are distinct phenomena, or indeed the only phenomena that compose our mental life. An opposing view, for example, is that mental life is nothing more than the manipulation of symbols. Further, these questions may invoke deeper questions regarding ontological commitments.

The content of this paper, then, may be summarized as follows: (i) a review of the essence of the state of the science in computational brain and mind studies, (ii) a series of questions regarding the inclusion of religious concepts in these studies, and (iii) the implications of such questions for both computer science and religious studies.

Computational Representations of Human Thinking and Religious Concepts

We wish to explore the relation between computational representations of human thinking and religious concepts [1]. The reader may have seen some of the recent Hollywood movies involving life-like robots such as "Artificial Intelligence" and "I, Robot." Indeed, these are entertaining movies, and they show robots as having autonomous thought, a sense of moral obligation, and a sense of self. One is tempted to say, especially an educated thinker, that those are just movies and in no way grounded in truth. While there is something legitimate about this intellectual attitude, one nonetheless should recall that similar things were said about various science fiction movies some fifty years ago, only to see some elements of those movies in reality today.

While the question of scientific realism in Hollywood movies is interesting and worthwhile, in this paper we wish to examine the issue of computational reasoning and religious concepts in a rather different methodological fashion. In the first part of the paper, we'll put Hollywood out of our minds and instead look at some of the basic reasoning paradigms that have issued out of academic artificial intelligence (AI) and robotics communities. Further, in this part of the paper we won't concern ourselves immediately with how religious beliefs, moral obligations, and the like might figure into those representations. That would be putting the cart before the horse. Instead, our goal is to understand simply the essence of various AI reasoning paradigms and how they constitute the mental life of a robot.

In the second part of the paper, we return to consider how various religious concepts and beliefs might be included in those computational representations. As an illustrative example, we'll consider primarily one such concept – that of moral/religious obligation. In particular, we'll look at the concept of "ought-ness" – e.g. I ought to give to charity, Tom ought to protect his brother, and the like. Let us point out that moral/religious obligation belongs to a higher category of religious concepts, and therefore we should be careful not to over-generalize from this one concept. Nonetheless, if we stay focused on this one concept, we may find it easier to apply our methodology to other related religious issues.

In the third and final the part of the paper, we'll consider the hard question of how computational representations of AI might affect religious studies and attitudes, and conversely, how religious studies and attitudes may affect our understanding and approaches to AI and robotics. There is precious little on this topic in the literature. Most of the AI research in the last fifty years or so has concentrated on understanding the reasoning processes of human beings and how to re-present those processes in a computer program. There have been a large number of successes where computer programs can perceive the world, identify objects, reason about the objects, and make decisions; further, these successes have been applied to real, practical problems. It is only recently that some AI researchers have begun to investigate how one might re-present human emotions and desires in a computer program. For example, a recent research paradigm is a belief-desire-intention (BDI) model of human reasoning, often couched in terms of a computational autonomous agent [2]. The gist of the approach is that a computational agent maintains a set of beliefs about some state of affairs, where a "desire" is defined as a goal of bringing about some other state of affairs, and an intention is defined as a series of actions to actually bring about the other state of affairs.

We can pose one question up front that will help set the context of our discussion as we proceed. Plainly put, we all know that we have desires, e.g. a want of something or an urge to perform some action. The BDI model, however, defines a desire as a goal of achieving some state of affairs. Now, the question is whether our desires, as we experience them, are the same as a state of affairs. I think we would all say "no." There is something like a "raw feel" when I experience a desire that is clearly missing in a description of a state of affairs with pen and paper, or in a computer program.

Note that the BDI model re-presents a desire in a computer program, where emphasis is placed on the hyphen between "re" and "presents." One might argue that we are doing our best to present the "raw feel" of a desire so that we can use it in a computer program. That's good, but therein is the difficulty also. It is logically possible that we start thinking that a desire is synonymous with a description of a state of affairs. This, then, is one way in which achievements in AI and robotics might influence our own understanding of ourselves, and it is worthwhile to consider implications of such possibilities ahead of time in case we think it a good idea to circumvent them.

The Mind and the Brain

Let us make an important distinction between the mind and the brain. This distinction is usually attributed to the French philosopher Rene Descartes in his *Meditations*, and the majority of scientists (but not all) have been guided by the distinction in their research on mind and brain to the present day [3, 4].

We'll approach the distinction with a thought experiment. Close your eyes and try to think of a cat. Some people are able to actually visualize a cat; some people will see the letters CAT; and other people will see nothing but blackness with perhaps a faint outline of a cat. In any case, we can all form some vague idea of a cat apart from actually perceiving one. Now the story goes as follows: a neuro-scientist can insert miniscule probes in one's head and measure electrical/chemical activity when one is thinking of a cat, or perhaps better when one has an urge of hunger or some strong desire. This is brain activity, and many experiments have been conducted where brain activity associated with some phenomenon such as hunger or love can be localized to a particular area of the brain. However, the neuro-scientist can not find, can not see, your idea of a cat. The scientist can open up the brain in search of the picture of a cat, but it won't be found. The vision of the cat, indeed the range of ideas and symbols in our heads, constitute mental activity.

Thus, the distinction between brain activity and mental (i.e. mind) activity is practically universally held in science. Typically we say that the brain is a physical thing than you can touch, stick probes into, and measure the activity of. The mind, on the other hand, is not a physical thing. We can't stick probes into it and we can't directly measure the activity of it.

The chart below shows some of the distinguishing characteristics of the mind and the brain:

Mind	Brain
Symbolic	Sub-symbolic
Mental	Physical
Unbodied	Bodied
Thinking	Unthinking
Unextended	Extended

The latter two lines of the chart were Descartes' views on the difference between the mind and the brain: The mind is a thinking and unextended substance, while the brain is an extended and unthinking substance. In fact, we recall from "Philosophy 101" that Descartes held these two kinds of substances to be the primary substances of our existence: Pick out any item in our experience and it will be a mental item or a physical item – there is nothing else. (As an aside, Descartes actually taught a third substance in the *Meditations*: a spiritual substance, or God, but interestingly we aren't taught that in Philosophy 101.)

Again, most but not all people, including scientists and people on the street, accept the distinction between the mind and the brain. The distinction can be controversial, but for our purposes let us accept it as at least an assumption of truth. As we'll see, the distinction has guided researchers in artificial intelligence and cognitive science in their research towards re-presenting human reasoning processes in computer programs.

There is a catch, however – an outstanding, unresolved problem. It would seem that there is some relation between the mind and the brain. That is, the brain would seem to be a vehicle for our mental experiences. One can imagine brain activity without a mental experience, but one would find it difficult to imagine a mental experience without brain activity. The catch, then, is to understand the precise relation between the brain and the mind. This bare view is usually referred to as epiphenomenalism – viz. the view that mental activity is a side-effect, or a function of, brain activity; however, an understanding of the precise connection between the two has been problematic.

Approximating Mind Activity with Rule-Based Reasoning Systems

Let us first consider how one might re-present mind activity in a computer program. A common approach to re-presenting mind activity is to couch knowledge in a rule-based reasoning (RBR) system [5]. See Figure 1. The basic structure of an RBR system consists of three parts:

- A Working Memory
- A Rule Base
- A Reasoning Algorithm

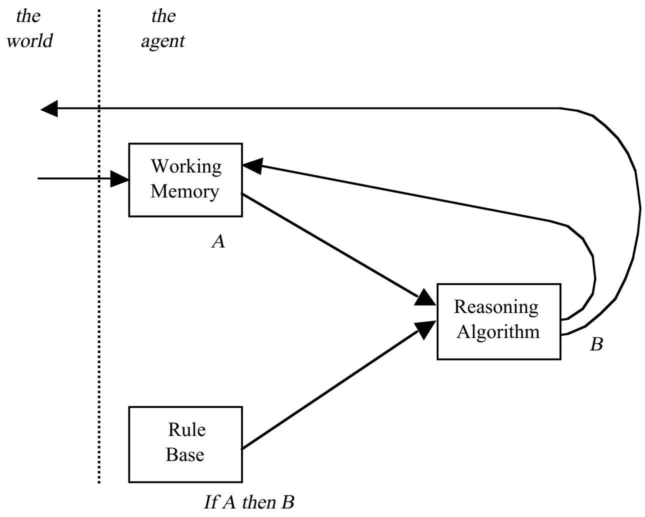

Figure 1. Mind Activity: The Basic Structure of RBR Systems

The working memory consists of facts. The rule base represents knowledge about what other facts to infer, or what actions to take, given the particular facts in working memory. The reasoning algorithm is the mechanism that actually makes the inference.

The best way to think about the operation of the reasoning algorithm is to recall the classic *Modus Ponens* inference rule in elementary logic:

A	*A fact in working memory*
If A then B	*A rule in the rule base*
Therefore, B	*An inference by a reasoning algorithm*

In this simple example, since the antecedent *A* of the rule *If A then B* matches fact *A* in the working memory, we say that the rule fires and the directive *B* is executed. Note that *B* can be several kinds of directive:

- Add a new fact to working memory.
- Perform a test on some part of the world and add the result to working memory.
- Query a database and add the result to working memory.
- Query an agent and add the result to working memory.
- Execute a control command on some world component
- Issue an alert.

The following is a simple illustration of how RBR systems are applied to traffic congestion on the Internet. Let us not worry about the technical details of the Internet; for our purposes we simply wish to gain insight into how RBR systems work. We can imagine how similar systems can be built for other domains of interest.

```
if      load(N1, t1) = high and
        packet_loss(N1, t1) = high and
        connection_failure(C3, S, t1) = true and
        connection_failure(C4, S, t1) = true
 then   add_to_memory(problem(t1) = congestion)

if      problem(t1) = congestion
then    add_to_memory(measure_traffic(t1–10, t1))
```

Regardless of the particular directive, after the reasoning algorithm makes a first pass over the working memory and the rule base, the working memory becomes enlarged with new facts. The enlargement might be a result of directives such as measuring traffic, or it might be a result of the agent entering new facts in working memory over time. In either case, on the second pass there might be other rules that fire and offer new directives and therefore new facts, and so on for each subsequent pass. At this juncture we should be able to appreciate the sort of complexity entailed by representing mental activity with RBR systems. Mental activity consists of continuous iterations over a fact base and a rule base.

Approximating Mind Activity with Case-Based Reasoning Systems

Let us now consider a second way that mind activity is re-presented in computer programs – Case-Based Reasoning (CBR). The basic idea of CBR is to recall, adapt, and execute episodes of former problem solving in an attempt to deal with a current problem. The approach is modeled on legal procedure: When an attorney represents a client allegedly accused of a wrong-doing, the attorney gathers as many facts and evidence surrounding the alleged crime as possible. With this information, the attorney pores over prior cases in case books in a law library. If the attorney is lucky enough to find a prior case which is similar to the client's case, the argument in defense of the client is as follows: Dear Judge and Jury: My client's case is almost exactly like the 1942 case of X vs. Y in which the outcome was not guilty. Accordingly, I propose that we save the court time and money by transferring the verdict of not guilty to my client's case.

Similarly, in CBR, former episodes of reasoning in some domain of interest are represented as cases in a case library. When confronted with a new problem, a CBR system retrieves a similar case and tries to adapt the case in an attempt to solve the outstanding problem. The experience with the proposed solution is embedded in the library for future reference. The basic structure of a CBR system is shown in Figure 2.

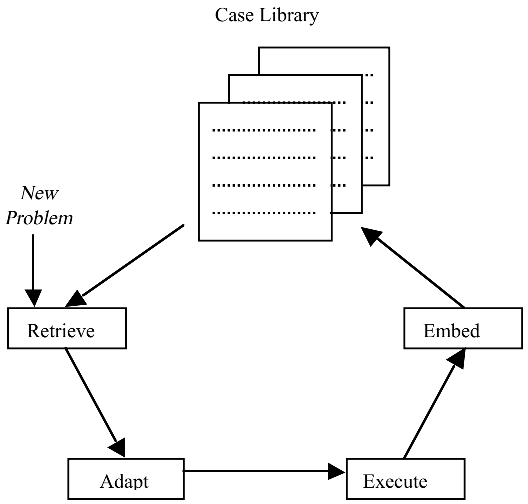

Figure 2. Mind Activity: The Basic Structure of CBR Systems

At this juncture we should have an essential understanding of re-presenting mental activity via CBR systems – our second paradigm. Let us now turn to re-presenting brain activity with neural networks.

Approximating Brain Activity with Neural Networks

Thus far we should notice that mind activity involves a knowledge and a manipulation of symbols. In the computer statement "device = server1," the name "server1" is a symbol. Of course, we assume that "server1" refers to some real physical object in the world. Neural networks (NNs), on the other hand are sub-symbolic – they don't reason with symbols such as "server1." The typical phrase that describes a NN is this: The knowledge is in the weights – as explained below [6].

NNs are modeled on what biologists know about the human brain. The brain consists of roughly 10 billion neurons, where a neuron may be connected to another neuron via a synapse. There are roughly 60 trillion synapses in a typical brain. Perception via the five senses may excite a subset of these neurons, which may cause the recognition of an object. These neurons, in turn, may excite other neurons which cause associations with the initial perception, e.g. consider the raw feel of "that object is a dog; and sometimes dogs bite strangers."

Figure 3 shows the basic structure of a NN. This is a quite simple NN for illustration purposes. We have 9 neurons $n_1 - n_9$ and 14 connections $w_1 - w_{14}$. Now, we want this NN to learn to associate an input vector of 1011 with an output vector 110. How does this happen?

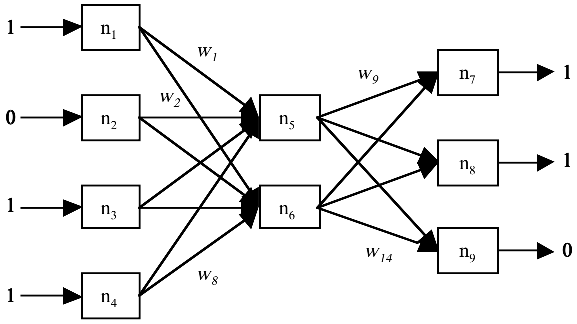

Figure 3. Brain Activity: The Basic Structure of a Neural Network

Each w will have a weight attached to it – a value between -1 and 1. In the beginning, these weights are generated randomly. An input vector 1011 is presented to the NN, whereupon some 3-bit output vector is generated as follows: for input to neuron n_5, calculate the sum of each input neuron multiplied by its connection weight, e.g. n_1 times w_1, plus n_2 times w_2, and so on. That sum will be the input for n_5. Now, n_5 contains a threshold function such as the following: if that sum is a positive number, then output a 1; if it is negative, then output a 0. In this way, the NN will ultimately output some 3-bit number, which likely will not be 110 because the initial weights were generated randomly. The NN looks at the difference between the actual output and the desired output 110, and then adjusts the weights accordingly so that the desired output is produced.

Let us imagine several pairs of 4-bit inputs and 3-bit outputs that we want the NN to learn. That means that the NN has to be presented with each pair and the weights adjusted accordingly as explained above. Typically, the NN has to be presented with hundreds of the same 4-bit/3-bit pairs before it can get the weights just right. This is what we mean by the expression "The knowledge is in the weights." The knowledge is the final weight vector $w_1 – w_{14}$, at which point the NN will have learned to associate the given set of input/output pairs.

This should be enough to explain the essence of NNs. Of course, NN research is quite more complicated than this. For example, we can imagine a 1000-bit input vector representing pixels in a picture where we want the NN to associate some input vectors with 110 and others with 010. Since the output is a 3-bit vector, we can manually correlate the output vectors with up to 8 different symbols. We could include an external table which says that 111 = cat, 110 = mouse, 100 = dog, 101 = bird, etc. Note, however, that the NN would not recall "dog" from an input vector; rather, it would recall 100 and via table lookup we find that 100 is a "dog."

Epiphenomenalism Revisited

We have discussed two ways to re-present mind activity in computer programs and one way to re-present brain activity. Let us now return to our assumption of epiphenomenalism – the view that mind activity is a side-effect or by-product of brain activity. Do we have anything like that in our discussion so far?

We note that the output of a NN, represented as some n-bit vector, may indeed be manually correlated with a symbol via some human-generated table. The connection, however, between say 100 and "dog" is not clear – at least this author is not satisfied with the simple table explanation. I don't think I have a table in my head such that my symbolic idea of a dog is associated with the activation of a subset of my neurons. There is no evidence of such.

We note that many "hybrid" reasoning systems have been developed that combine a NN with an RBR or CBR system, where the NN recognizes things in the world, the table associates those things with symbols such as "dog," and then a RBR system applies a rule such as "If X is a dog then X might bite." Many practical reasoning systems have been built using this sort of hybrid architecture. However, we point out that the unclear piece of the architecture is the table that associates 100 with "dog." Perhaps we should think of the table metaphor as a place-holder for further research on a hard, but heretofore unanswered question.

Representations of the Brain and the Mind in Computer Science: Is there Room for Religious Concepts?

We are now in fair shape to consider the question in the title of this paper. As indicated in the introduction, we won't consider the entire array of religious concepts, although that might be fodder for further thinking and analysis. Let us consider only the concept of religious/moral obligation as an illustration [7]. Here is the kind of dialogue we might encounter:

D1: Of course there is room for religious concepts in computer programs and robotics because it has already been done. There are computer programs that include a symbol for "God." There are computer programs that include a symbol for "the Good." There are computer programs that include symbols for obligation, e.g. "ought(X)" where X is some action or state of affairs. There are even computer programs that infer obligations by applying the 10 Commandments, which are rules and therefore quite appropriate for re-presentation in a computer program.

D2: No. The symbol for "ought" does not capture the human raw feel of obligation. It is barren, without life, and a poor substitute for our human experiences of obligation. If we have a robot whose program provided an output of "ought(X)" it simply won't have the human raw feel of ought-ness. It isn't possible.

D1: Hold on just a minute. Let's take the distinction between mind and brain as granted. Biologists can measure brain activity as a stimulation of localized neurons, where the measurement is in the form of electrical/chemical activity. That we know. Now, suppose we burned a neural network on a chip, so to speak. Surely, we can also measure stimulation in the chip in terms of electrical activity. After all, the bits, the 1s and 0s in a NN, are at bottom represented as electrical pulses in computer hardware. Now, if we can assume that the electrical/chemical activity in the human brain is at least analogous to electrical impulses in a computer chip, then we do indeed have something like the raw feel of obligation that you say is missing in computer programs. Better, if science has uncovered some substance or theory that incorporates both electrical activity and brain activity, then we have further evidence that a robot's brain can experience a raw feel.

D2: I'll accept that for the time being; however, you'll have to prove that an electrical pulse of a computer chip is sufficiently analogous to human brain activity. For the sake of argument, we'll say it's a simplifying assumption only, but I don't accept it outright. Now, would you say that the feeling of obligation, or of religious sentiment in general, rightly belongs to one's brain activity or one's mind activity? It would seem to belong to brain activity. If we accept that, then we should try to explain how religious sentiment plays into our mental life. I don't think we want to say that religious sentiments are purely and only brain-related.

D1: Agreed. That takes us back to the problem of the gap between brain activity and mind activity – the so-called table problem. That's a difficult challenge indeed.

This sort of dialogue may well take a number of twists and turns as we try to speak to the topic of this paper. In the next and final section, allow me to go out on a limb to formulate some opinions on the matter.

Conclusions

1. Some people argue that we have more important things to think about than whether religious sentiments can be expressed in computer programs, or whether robots can or should be able to ponder and develop religious attitudes. I agree with part of that statement – there are indeed more important things to think about in the present. However, I disagree if the arguer is implying that we shouldn't think about these questions at all. In fact, I would counter that those questions are extremely important and that we should be thinking about them now. One only has to consider that 200 years ago it would have been appropriate to have started thinking about the religious, social, and ethical implications of planes, trains, automobiles, and telephones. Further, 10 years ago it would have been, and was, appropriate to start thinking about the possible ramifications of the Internet and global communication. At this moment, it is appropriate to think about the ramifications of a global, wireless Internet wherein individuals can communicate via voice and video in near real time. So, let those of us who are passionate about the subject matter continue doing so. It is worthwhile. There are plenty enough problems to go around.

2. There is room for religious sentiments and concepts in computer science – not only in virtue of their existence already, but in their capacity as being guides for understanding intelligent systems for the sake of both intellectual curiosity and practical application. In moral reasoning, deontological theories, i.e. those theories that hold that rational/moral action is grounded in some given set of rules such as the 10 commandments or Kant's categorical imperative, provide insight into how to build intelligent systems. Likewise, teleological theories such as Utilitarianism, i.e. theories holding that right action is decided by the greatest good for the greatest number, also provide insight into understanding intelligent systems. Conversely, computer science can provide insight into certain areas of religious studies. We considered one problem in the previous section in which we argued that there isn't a clear understanding of the relation between the brain and the mind. Once one brings down theories about the mental life of humans to the computational

level, there is at least the possibility that missing pieces will be uncovered. We argued above, for example, that a table that correlates artificial brain activity with artificial symbols (e.g. 100 = dog) is simply a place-holder for something we haven't figured out. One philosopher might put it: It is something, I know not what. Another philosopher might put it: There is a cause and effect, but I cannot find the connection between them.

3. I want to believe that it is in theory possible for future robots to experience something like what we would call a "raw feel" of religious sentiment, but it is difficult to believe it. It depends in part on what our dialoguer *D1* argued above: Whether we can say that electrical activity in a computer chip is sufficiently analogous to electrical/chemical activity in the human brain, or whether there is some underlying substance or mechanism that is common to both. We should keep in mind, though, that the measurement of something is not the something itself. However, if we push that idea to its limit, we force ourselves into solipsism, or the view that my immediate mental life is my only reality, which most everybody holds as untenable. Further, the possibility of future robots being capable of experiencing religious sentiments depends in part on the problem addressed in #2 above.

4. Finally, in closing, I would like to comment on the art of deception. Clearly, people don't like to be deceived. Here are few personal eccentricities. I can't go into the lobby of a high-class hotel without determining whether a beautiful plant is real or artificial. Sometimes I have to go touch it. I can't see a shiny souped-up Roadster on the street without trying to determine whether it is really a VW chassis with a fabricated Roadster body. I visited my mother once and noticed a cat curled up sound asleep beneath her desk, but after about 30 minutes it occurred to me that the cat hadn't budged. It was a fake cat, and I highly disapproved of my Mother having a fake cat in her study. I was practically offended. We don't have to mention spam and junk email. In sum, I'm pretty sure that humans don't like being deceived. I don't think that people have to worry too much about being deceived by a robot – i.e. being deceived into thinking that the robot is a real human being.

References

[1] Lewis, Lundy and Ted Metzler. Artificial Intelligence and Robotics: Implications for Theology and the Religious Community. Forthcoming as a Video Presentation, Good Star (Growing Open Oklahoma Dialogue in Science, Technology, and Religion), http://starport.okcu.edu/SI/GS/. 2005.

[2] Rao, A and Georgeff, M. BDI Agents: From Theory to Practice. In Proceedings of the *First International Conference on Multiagent Systems* (ICMAS'95). 1995.

[3] Churchland, Particia. *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. MIT Press, Cambridge. 1989.

[4] McCorduck, Pamela. *Machines Who Think*. W. H. Freeman and Company, San Francisco. 1979.

[5] Lewis, Lundy. *Managing Business and Service Networks*. Kluwer Academic/Plenum Press. 2001.

[6] Negnevitsky, Michael. *Artificial Intelligence*. Addison-Wesley. 2005 (2002).

[7] Castaneda, Hector-Neri. *The Structure of Morality*, Charles C. Thoman: Springfield, Illinois. 1974.