

Advances In Consciousness and Artificial Intelligence, with Questions About Their Possible Implications for Theology and Philosophy

by Suzanne M. Sharrock,
Department of Quantitative Methods and Computer Science

1. Introduction: Motivation and Overview

I started this seminar project by trying to choose some areas within science, including at least one from my field of computer science, on which to focus my reading and thoughts. I settled on two areas of strong interest to me, that I believed had some relevance to theological/philosophical discussion. Within computer science, the most obvious area of connection seemed to be artificial intelligence (AI). Computing machines now have some considerable ability to perform reasoning; this causes a number of people to say such machines already exhibit “intelligence” in a variety of domains. This is a faculty that many still believe only living creatures possess. Furthermore, it is often attributed to human creatures only, possibly for theological and/or philosophical reasons (e.g. we are the culmination of God’s creation with the highest intellectual and cognitive abilities).

In addition to the area of computer science, I am also intrigued by current discoveries and advances in the study of consciousness: what it is, the mechanisms behind it, what we are learning about it, what we might not be able to find chemical/biological mechanisms for — these are all topics I wished to explore. Consciousness, to many, is at or near the core of what it means to be human. Understanding consciousness, at least as many aspects of it as we can, seems a fairly noble human endeavor. Paul Davies’s words express this better than I: “A universe in which the emergence of life and consciousness is seen, not as a freak set of events, but fundamental to its lawlike workings, is a universe that can truly be called our home.”[DAVI92]

So I set out to do some reading and reflection within these two distinct, but slightly related areas. For example, if we fully or mostly understand consciousness, and consciousness can be fully explained in terms of physical laws, it is likely that we can build a machine to act as a conscious creation acts. The field of AI (as well as numerous other areas of computer science — architecture, networks, programming languages, graphics, etc.), which even now must deal with tremendous amounts of data and computation, as well as synthesis of data, would be further challenged. Whether or not that machine then might be considered “conscious”, or would have any relevance to a philosopher or theologian’s concerns, is not a question I can answer. However, the prospect is intriguing, if somewhat speculative at this point.

I have chosen to concentrate in this paper primarily on the AI and consciousness issues, and bring in questions and ideas from the readings in philosophy and theology as they

related to the scientific material. I have done this primarily at the level I would use in introducing my students to some current research issues in science, and their relationship to the “bigger questions”. Advances both in AI and in consciousness do have ties to theological and philosophical ideas about human nature and purpose, divine creation, and divine purpose. I hope that our discussions can give me additional insights into relations between the areas.

The remainder of this discussion paper is divided roughly into three distinct sections, followed by a conclusion. The following section (Section 2) is a list of questions that illustrates how I view some overlapping concerns among the fields of consciousness, AI, theology, and philosophy. The next section (Section 3) is primarily a synthesis of some aspects of current consciousness research, and the thoughts and opinions of those working in the field. The last section (Section 4) describes those areas of AI that relate closely to the idea of machine intelligence — beginning with the quest for a computer that could think (the Turing Test) and moving to the current state of AI research, particularly in expert systems. I’ve included the thoughts of several prominent researchers in those areas. In the concluding section, I try to give a sense of the optimism and wonder I myself feel over advances within these fields, even while realizing that these advances may raise, more than answer, questions about the human person, and therefore eventually, about “truth”, “faith”, and ultimate

2. Questions

I would like to start with some general questions that motivated me to look more deeply into these areas. Is consciousness one of those properties at the core of “being human”? Is it at the core of being “special” in God’s creation? Which aspects of consciousness are unique to humans? Can we ever really know? How much of consciousness, if any, is special (i.e., not physically predictable), or is it all just the natural byproduct of a complex brain, emerging like wind from intricate weather patterns? How will we know when we know? Does predictable mean not special? Does the “specialness” of consciousness really matter to a purposeful life? What about hope, love, benevolence? Are some of these “transcendental phenomena” (i.e. in the sense of Berger’s definition: “phenomena that are to be found within the domain of our ‘natural’ reality but that appear to point beyond that reality”[MCBR81, p. 193])? They certainly seem to be part of consciousness. Are (any or all of) these parts uniquely human?

Would a biological/physical complete understanding of consciousness help/hurt the philosophical/theological understanding of humanness? What does it mean for a machine to be “conscious”? Is there even any point in trying to imagine a conscious machine? Would this cause any ripples in the

As a slight aside, what does it mean theologically if we are able to use genetic engineering to create humans with only specifically chosen (by humans) characteristics? Does the method of human life beginning (deliberate, by human engineering, or otherwise) contradict (Christian and other) religious belief?

How intimately are intelligence and consciousness linked? Depending on the answer to the previous question, what degrees of intelligence exist independently or semi-independently of “consciousness”?

3. Thoughts and Readings on Consciousness

(My readings on consciousness have come from the conference “Toward a Science of Consciousness,” Tucson, Ariz, April 1996 [ANDE96, BLAK96], as well as from the lectures of Henry James [JAME39], and Richard McBrien’s Catholicism [MCBR81].)

I’d like to start with what I would call the “highest” view of consciousness. Basically, it is that we not only know, but “we know that we know. The more conscious we are of ourselves, of our knowing powers, ... of the implications of our thoughts, our judgments, and our actions, the more we are in possession of ourselves.” [MCBR81, pp. 148-149]. We might say that the more we learn about consciousness, the more we truly know about humans. So results in this area can be of use to scholars in both the physical sciences and in areas of philosophy and theology.

As I understand it, the two main areas of current research into consciousness are concerned with what are called the hard problem (or problems) and the soft problems of consciousness. Although I will describe each of them in more detail below, briefly we may think of the hard problem as addressing the “why” of consciousness, and the soft problems as addressing the “how”; however, these are not absolute distinctions.

3.1 The Hard Problem of Consciousness

The “hard problem” of consciousness can be described by a series of questions. “What is the nature of subjective experience? Why do we have vividly-felt experiences of the world?” [BLAK96] At the Tucson conference, Dr. Chalmers of UCSC stated it in the following way: “When we see, we experience visual sensations — the felt quality of redness, the experience of dark and light, the quality of depth in a visual field. Other experiences go along with perception in different modalities — the sound of a clarinet, the smell of mothballs. ... Then there are bodily sensations from pains to orgasms — mental images that are conjured up internally, the felt quality of emotion, and the experience of a stream of conscious thought. What unites all of these states is that there is something it is to be in them. All of them are states of experience.” [ANDE96] People such as he argue that consciousness “has absolutely unique properties: it is private, subjective, peculiar to the individual, and cannot be directly observed by a third person.” [ANDE96]

Pat Churchland (UCSD) noted that “many people ... want to believe in a soul, life after death and the specialness of humans and their inner thoughts. They have a negative gut reaction to the idea that neurons — cells that can be probed under a microscope — are the source of the ‘me-ness of me’”. [BLAK96]

It has been suggested that consciousness might turn out to be an irreducible property, in the same category as time and space.

Again, Chalmers is one holding those views: “My approach is to think of conscious experience itself as a fundamental property of the universe. Thus the world has two kinds of information, one physical, one experiential. The challenge is to make theoretical connections between physical processes and conscious experience.” [BLAK96]

A more radical (and not well received!) view proposed by Roger Penrose and others involves switching between quantum mechanical and classical states inside certain brain proteins to produce our conscious experiences, that over time, give rise to conscious thought. In one of the criticisms of this view, Dr. Pat Churchland stated: “Their logic is, consciousness is deeply mysterious, quantum mechanics is deeply mysterious, ergo the two are the same mystery.” [BLAK96]

Yet others, including McGinn from Rutgers (a philosopher) believe that for humans to even begin to grasp how subjective experience arises from matter “is like slugs trying to do Freudian psychoanalysis — they just don’t have the conceptual equipment.” [BLAK96] Dr. Forman (Hunter College, religion) believes that mystical experience has something to tell people about consciousness. “To understand genes we look at bacteria like E. coli. To study memory, we analyze the memory of a sea slug. But to probe consciousness, we need to examine the experience of mystics, who experience their own consciousness in its simplest form.” [BLAK96] Furthermore, he stated that millions of people regard these types of experiences, feeling a one-ness with the universe, as the highest experience that the conscious brain has to offer.

In summary, while most agree that the “hard” problem is hard, there is considerable disagreement about if and how it will ever be solved. The soft problems, while numerous, do not give rise to so much disagreement about whether solutions can be discovered.

3.2 The Soft Problems of Consciousness

The “soft problems” of consciousness include such questions as: How does sensory information become integrated in the brain? How do we see and reach out for an object? How are we able to verbalize our internal states and report what we are doing or feeling?

As summarized in [ANDE96], for several centuries our view on consciousness was pretty much summed up by Descartes: “When I consider the mind, that is myself in so far as I am merely a conscious being. I can distinguish no part within myself. I understand myself to be a single and complete thing. Nor can the faculties of feeling, will, understanding and so on be called its parts, for it is one and the same mind that will, feels, and understands.” Furthermore, it was believed that consciousness was/is a uniquely human characteristic.

However, today neurologists disagree with Descartes, no longer viewing consciousness as an “all or nothing” phenomenon. In particular, there is strong evidence that various

processes which form parts of consciousness are assembled from different, distinct neuronal processes in the brain. Two particular examples cited are from studies on brain-injured persons. The first example is from studies with “blindsighted” people who have damaged portions of their visual cortex. The second is from people with a particular type of stroke damage called asonognasia, where the right side of the brain has been injured.

Blindsight is at odds with our normal, common-sense view of consciousness. Here, people who have lost all sensation of light and/or color from an area of their visual field (hence, respond “No” when asked whether they can see a particular object within that field) can nevertheless usually point quite accurately at where the object is within that “unseen” visual field! While they have lost all “conscious” sensation of seeing, at some level they are still able to “see”. The scientific explanation is that the visual pathway splits into many parallel streams as it approaches the primary visual cortex, with some streams going on to other areas of the brain to provide “unconscious” knowledge about the object. So blindsight offers clues as to which parts of the brain are used to generate various attributes of visual consciousness.

Asonognasia, on the other hand, is caused by stroke damage to the right side of the brain, which paralyzes the left side of a person’s body. Despite the paralysis, and in the face of obvious evidence to the contrary, anosognosics claim (and apparently believe) that their paralyzed side works perfectly well! So, for example, people insist they are performing a task (such as clapping with two hands, or touching an object with the paralyzed arm/hand) even when they can “see” that this is not the case. In fact, we would say, they cannot “see” it as we do, even though there is nothing else wrong with their minds. It does not seem to be a psychological delusion, and shows up exclusively in people with left-side paralysis, not right-side paralysis. The theory proposed by Ramachandran is that “Anosognosia is a problem of the mind’s belief system, not its perceptual system. The mind needs a theory of the world in order to organize and make sense of the constant stream of sensory inputs. But the theory-making part of the brain must also be able to ignore inputs that don’t fit with its world view, lest every mistaken

The right half of the brain acts as a devil’s advocate. When too much conflicting data accumulates — for example, repeated awareness that the left arm cannot move — the devil’s advocate overcomes the left brain’s defense mechanisms and forces it to restructure its world view to fit the new information. In people with anosognosia that mechanism — your devil’s advocate — is damaged, and the left brain is free to pursue a strategy of denial and confabulation. There is no limit to the delusion.

The above are two of many examples of attempts to solve these “soft problems”. And to relate back to the hard problem, some of the researchers believe that when enough of the “soft problems” of consciousness have been solved the hard problem will also be mostly or completely solved. A proponent of this view, Daniel Dennett (Tufts) stated it this way: “Mental states do not become conscious by entering some special chamber in the brain, nor by being transduced into some privileged and mysterious medium but by winning the competition against other mental states for domination in the control of behavior. No more is needed. Consciousness is an epiphenomenon.” [BLAK96] Those who think that

brain processes cannot explain our first-person experience of consciousness have the question all wrong, according to him. “It presupposes that what you are is something else — in addition to all of this brain-body activity. But what you are, however, just is the organization of all this competitive activity between this host of competencies which your body has developed. You automatically know about these things going on inside your body because if you didn’t it wouldn’t be

4. AI and its Relationship to “Human Intelligence”

As many of us have already stated and believe, science, mathematics, and many disciplines work within frameworks of assumptions, some of which are not usually articulated. These constitute what we might call the faith of our science. The field of artificial intelligence also has its “faith.” Feigenbaum, a renowned AI researcher, stated: “Since Turing, the faith of AI has been that human intelligence is best modeled as software for a physical symbol system — the digital computer.” [FEIG96], although not all practitioners and theorists in the field would put it quite so strongly.

While we might start with many different possible definitions of intelligence, one that seems quite relevant here is given by Allan Newell. “Intelligence is the ability to bring to bear all the knowledge that one has in the service of one’s goals. To describe a system at the knowledge level is to presume that it will use the knowledge it has to reach its goal. Pure knowledge-level creatures cannot be graded by intelligence — they do what they know and they can do no more, that is, no better. But real creatures have difficulties bringing all their knowledge to bear, and intelligence describes how well they can do that.” [NEWE90]

With this as a starting definition of intelligence, and keeping in mind the purpose of our seminar, we might also ponder the particularly interesting question: “Might we be able to construct a machine ‘more intelligent’ than we, that could give us a ‘proof’ of the existence of God — as software has helped prove the mathematical Four-Color Theorem?” To some it may seem either naive or arrogant to even propose such an idea. I will revisit this

Shifting back to the AI area itself, I would like to put forth various views of what machine intelligence is, including the historical Turing test, and discuss a little about where we are with some current levels of machine intelligence and what some of the current difficult problems are in this area.

4.1 Beginnings of AI: Turing and His Test

Alan Turing, logician and theorist, cryptographer and computer scientist, is considered one of the true “fathers” of artificial intelligence. Turing also had a deeply “applied” view of design, preferring an operational definition of machine intelligence/thinking to simply theorizing about the characteristics of an intelligent machine. In his famous 1950 paper, “Computing Machinery and Intelligence,” he proposed that we could decide if a machine was intelligent if it could pass what is now called the “Turing Test.”

The Turing Test is both historically important at the beginning of the search for machine intelligence, and a subject of continued interest and software research and development in the 1990s. While many now view the ability of a computer to pass this test as insufficient grounds for machine thinking or machine intelligence, it served as an early challenge to be met. Because of its importance, I will first describe the test itself, as well as early and recent modifications to the test, then some fairly recent results with the test, and what those results mean (or don't mean). This includes challenges to the idea that passing this test shows machine intelligence.

Back to the original test. Turing proposed using an "imitation game" to determine whether machines can think. He described it as follows.

It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." [TURI50]

Turing then proposed replacing A by a computer and playing the imitation game. He argued that one could ask whether the machine could fool the interrogator, and if it could, that this would mean that the machine could think. The test has since then been commonly understood to be an imitation game involving a human and a computer (rather than a man and a woman), where the computer is imitating a human and the human is making no particular effort to imitate a computer. More recently, a group of computer scientists, philosophers, and a historian of science interested in computing came up with an n-way Turing test, in which judges deemed to be "average interrogators" (i.e. without particular computer expertise) would interact with a number of computer terminals, some connected to computer programs only and some connected to terminals operated by people. The judges would rank the terminals on how "human-like" the interchanges were, and on the basis of this ranking, decide which of these were humans and which were computer programs. (Conversations were restricted to be of an "everyday" sort.) The first test was held in 1991, and one of the entered computer programs fooled half (5 of 10) judges. Of course, this does not mean to most of us that computers actually "think" — the problem domain was restricted, and the program that won actually was designed to simulate human typing errors, so that some have said that the winning program actually exhibited "Artificial Stupidity" rather than artificial intelligence. [ECON92] Other cited flaws included the presumption that intelligent agents are equivalent, that human intelligence is highest (as opposed to say, an omniscient alien's intelligence), and that the test is testing humanity, not intelligence. [JOHN92],[FOST93]

Nevertheless, the concept of some type of test is still valuable. One commentator stated it this way: "how well one can fool someone is not a measure of scientific progress. The Turing Test is an empirical criterion: It sets AI's empirical goal to be to generate human scale performance capacity. This goal will be met when the candidate's performance is totally

4.2 Measuring Machine Intelligence: Beyond the Turing Test

So we are now back to more questions. If the Turing Test is itself not a good measure of machine intelligence, what is? In particular, what constitutes an intelligent system? Furthermore, can AI “equal” human intelligence? Can AI exceed human intelligence? I will give a few possible ideas from “experts,” with plenty of room for other ideas and disagreements; however I find these compelling.

One of the AI pioneers combines an operational definition of intelligence with a defense of it, and criticism of those who would question “intelligence” in machines:

I know of only one operational meaning for “intelligence”. A (mental) act or series of acts is intelligent if it accomplishes something that, if accomplished by a human being, would be called intelligent. I know my friend is intelligent because he plays pretty good chess (can keep a car on the road, can diagnose symptoms of a disease, can solve the problem of the missionaries and cannibals, etc.). ... The trouble with those people who think that computer intelligence is in the future is that they have never done serious research on human intelligence. Shall we write a book on “What Humans Can’t Do”? ... Computer intelligence has been a fact at least since 1956, when the Logic Theory machine found a proof that was better than the one found by Whitehead and Russell. ... Let’s stop using the future tense when talking about computer intelligence. (Newell, quoted by Reddy in [REDD96])

In other words, because programs can accomplish all of these goals (not necessarily in a single set of programs!), machine intelligence is demonstrated.

What others might argue is that the “intelligence” of such a machine is akin to that of an idiot savant — overwhelmingly good in its specialized areas, but in no others. Arguments might also be raised linking a lack of “consciousness” of the machine to its lack of potential for true “intelligence” (something that the friend mentioned above certainly has).

Nevertheless, there is something very practical and positive about machines that can “think,” even if only within certain specified domains. Feigenbaum and others at Stanford believe that “the means for intelligent action is primarily knowledge, and in most practical situations domain-specific knowledge. ... People are more than broad generalists with weak reasoning skills. Their most important intellectual behaviors — from the points of view of our culture, technology, and economy — are the behaviors of expertise, of specialists deeply trained and knowledgeable about the specifics of their various domains of work.”[FEIG96] In many cases, an expert-system (knowledge-based AI program) is able to more reliably predict or diagnose problems than a human expert could, because of its (potentially much) larger knowledge base within the specific area, and humans’ computational constraints.

This might lead us to one more question — “Can AI equal human intelligence?” (Although some AI experts would turn the question around!) The seemingly reasonable

answer to AI expert Raj Reddy, as well as to myself, is: “AI can be both more and less than human intelligence. It doesn’t take large tomes to prove that they cannot be 100 percent equivalent. Ultimately, what will be accomplished by AI will depend more on what society needs and where AI may have a comparative advantage than on philosophical considerations.” [REDD96]

Reddy goes on to talk about these differences by means of analogies, including a “book” in a digital library, which is both more and less useful than “real” book. For example, you cannot use the digital book as part of your rare book collection, nor can you light a fire with its pages, nor can you throw it at someone without incurring great expense! However, you can search much more quickly for many types of information within it, change font size if you are farsighted, etc. The point is that the electronic book is NOT the same as the real book, but both more and less.

One of the ways in which AI is “less intelligent” involves the unsolved problem of constructing “reasonable” algorithms that “know when they do not know.” A somewhat humorous example from John McCarthy is the following:

Suppose one asks the question, “Is Reagan (fill in current president here) sitting or standing right now?” A system with a large database of facts might proceed to systematically search the terabytes of data before finally coming to the conclusion that it does not know the answer. A person faced with the same problem would immediately say, “I don’t know,” and might even say, “and I don’t care.” [REDD96]

Some would argue “Isn’t AI just a special class of algorithms, written by a person or persons — nothing more, nothing less — completely limited by the ability of the human designer?” In a general sense, yes, AI is “just software.” However there are classes of AI algorithms that are becoming better and better at performing adaptive, goal-oriented behavior, learning from experience, integrating vast amounts of knowledge to perform a task, reasoning, etc. — (one particular area that comes to mind is neural networks) — and there are programs that write programs! So, I’ve tried to present a number of points of view of what constitutes machine intelligence, although I have just barely skimmed the surface. For example, there are large-scale adaptive communicating sets of expert systems (the CYC project at UT Austin [LENA92]) that attempt to give programs “common sense” and the ability to learn information about different domains from other expert systems. What I have tried to emphasize is that, to a certain extent, machines can and do exhibit many capabilities that not so long ago would have been deemed solely the province of humans.

But ultimately, one might completely question the purpose or usefulness of creating machines that duplicate human intelligence. Machine intelligence that duplicates human intelligence is not economical since “people are in plentiful supply [and] should a shortage arise, there are proven and popular methods for making more.” [ECON92] Perhaps a much more optimistic (and hopeful) motivation is “One day researchers may use the precision and power of computers to re-create human reasoning. In the process

they may unravel many mysteries — including, possibly, the roots of human intelligence.”

5. Conclusion

The advances and continuing progress of science to understand the mechanisms of human consciousness and to design and create machines that have many aspects of human intelligence (and ultimately, perhaps of human consciousness) can be viewed as either a very positive and enriching advancement of human awareness and human abilities or as a somewhat disturbing trend that upsets some traditional views of what it is to be human, particularly from a philosophical, psychological, or theological point. Perhaps our uniqueness, if that is what it is, is not quite of the form we once thought. But we have no absolute conclusions at this point. However, I take the optimistic view that every advancement we make in some way furthers humans’ special place in our universe, even if those advancements seem to lead us down paths whose end we cannot yet fathom. That is part of the wonder and mystery (insofar as we are “conscious” of wonder and mystery!) of our human life.

I would like to conclude on a stirringly optimistic note, with one more quotation, about AI, that expresses this belief so well. “The real challenge (of AI), then, is not to recreate people but to recognize the uniqueness of machine intelligence, and learn to work with it. Surrendering the human monopoly on intelligence will be confusing and painful. But there will be large consolations. Working together, man and machine should be able to do things that neither can do separately. And as they share intelligence, humans may come to a deeper understanding of themselves. Perhaps nothing other than human intelligence — constantly struggling to recreate itself despite crumbling memories and helter-skelter reasoning — could even conceive of something as illogical and wonderful as machines that think, let alone build them and learn to live with them.” [ECON92]

Cited References:

[ANDE96] Anderson, A., Holmes, B., and Else, L., “Zombies Dolphins and Blindsight”, *New Scientist*, May 4, 1996, pp. 20-27.

[BLAK96] Blakesless, S., “The Conscious Mind Is Still Baffling to Experts of All Stripes”, *New York Times Science*, April 16, 1996.

[DAVI95] Davies, P., “Physics and the Mind of God”, *First Things*, Aug./Sept. 1995, pp. 31-35.

[ECON92] (editorial), “Artificial Stupidity”, *The Economist*, Vol. 324, No. 7770, August 1, 1992.

[FEIG96] Feigenbaum, E., “How the “What” Becomes the “How””, *CACM*, May 1996, pp. 97-104.

[FOST93] Fostel, G., "The Turing Test is For the Birds", SIGART Bulletin, Vol. 4, No. 1, 1993, p. 7.

[HARN92] Harnad, S., "The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion", SIGART Bulletin, Vol. 3, No. 4, 1992, p. 9.

[JAME36] James, H., The Varieties of Religious Experience, Random House, NY, NY, 1936.

[JOHN92] Johnson, W.L., "Needed: A New Test of Intelligence", SIGART Bulletin, Vol. 3, No. 4, 1992, p. 7.

[LENA95] Lenat, D.B., "Artificial Intelligence", Sci. Am., Sept. 1995, pp. 80-82.

[MCBR81] McBrien, R. P., Catholicism, Winston Press, Mpls, MN, 1981, (in particular, p. 149 on consciousness).

[NEWE90] Unified Theories of Cognition, Harvard University Press, 1990.

[REDD96] Reddy, R., "To Dream the Possible Dream", CACM, May 1996, pp. 105-112.

[RUSS57] Russell, B., Why I Am Not A Christian, Simon and Schuster, NY, NY, 1957.