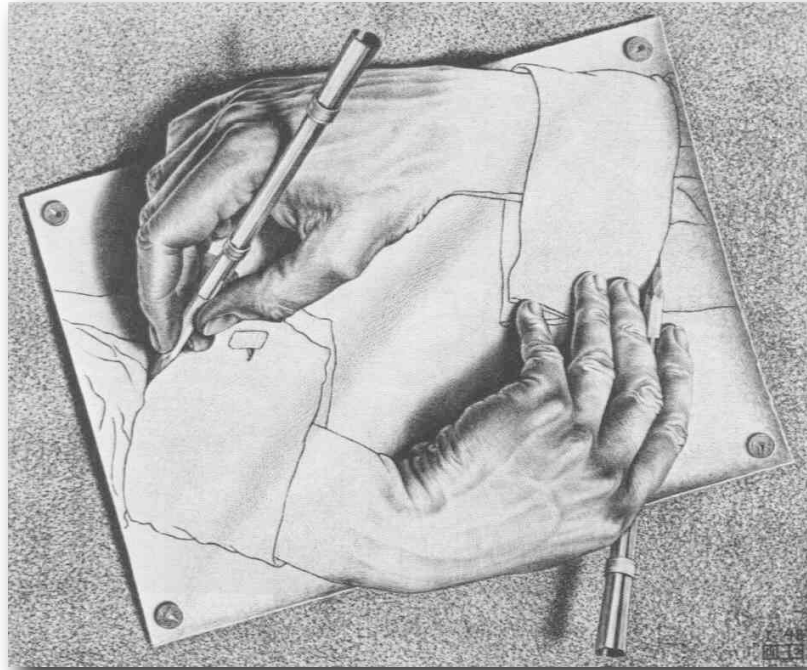


Creating Consciousness



M.C. Escher's *Drawing Hands* (1948, Lithography)

Martijn van Steenbergem
Friday, July 28th, 2006
Le Thoronet, France

For the course Philosophy of computer science
Taught by prof. dr. John-Jules Meyer and dr. Richard Starmans
At Utrecht University, the Netherlands

Table of contents

Preface.....	3
Introduction.....	4
What is considered creating?.....	4
What is consciousness?.....	6
Tests of consciousness.....	7
Where is the consciousness?.....	8
Reductionism.....	10
Emergent complexity.....	11
Reductionism: a vicious circle?.....	13
Objections.....	15
The Chinese Room.....	15
Lack of nonlocality.....	15
Continuous versus discrete.....	16
The limited brain.....	16
Other souls.....	17
Ethics.....	19
Conclusion.....	20
Bibliography.....	21
In The Mind's I.....	21
In Wikipedia.....	21
Meta.....	22

Preface

While writing this paper, I have made extensive use of a 1981 book called *The Mind's I*, a collection of stories and articles selected by Daniel Dennett and Douglas Hofstadter. They believe that it is theoretically possible for a computer to have consciousness. That is, there is no theoretical, abstract reason for it to be impossible, contrary to, for example, John Searle's ideas.

Each selection in the book is accompanied by commentary ('Reflections') from either Douglas Hofstadter, Daniel Dennett, or both. The selections include the famous pieces *Computing Machinery and Intelligence* by Alan Turing and *Minds, Brains, and Programs* by John Searle, as well as some provoking thought experiments such as *The Soul of the Mark III Beast* by Terrel Miedaner and *The Story of a Brain* by Arnold Zuboff.

Besides this (in my humble opinion) rich and varied source of ideas, I have also looked at both some newer and older ideas: quantum mechanics and Zen ideas respectively. Please see the chapter *Bibliography* for a complete overview.

Introduction

The idea of creating consciousness has become very popular in the past century. The quick development of various sciences, most notably that of computer science and biology, has impressed people and is the source of many philosophical arguments and thought experiments, as well as ideas for movies such as *A.I.* and *Bicentennial Man*. Will computers ever be just as witty, creative and emotional as humans? Can we *create consciousness*? That is the question central to this paper. I'll show you, the reader, various opinions on this question, and hope to stimulate your creativity.

There are differences between consciousness, intelligence and souls, but I will assume that for something to be conscious, it will have to have at least some intelligence.

Writing software is an obvious choice when trying to build a conscious being: computers are what is called "universal": when given enough time, they can, in theory, compute most things. As John von Neumann said, "You insist that there is something a machine cannot do. If you will tell me *precisely* what it is that a machine cannot do, then I can always make a machine which will do just that!"¹ The idea of intelligent or conscious machines is quite old. For example, in the earlier 19th century Charles Babbage and Lady Ada Lovelace worked on a difference engine, an idea originally from 1786.

John Searle in his 1980 article *Minds, Brains and Programs* distinguishes between two forms of artificial intelligence (AI): strong AI and weak AI. He writes:

According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states.

This paper discusses the strong version of AI. Before we can delve deeper into the matter of creating consciousness, let's first think about what we consider creating, and what we consider consciousness. While the definition of 'creating' is described in the next section, the elaborate topic on consciousness deserves a chapter on its own.

What is considered creating?

One answer to the posed question might be: we are already able to create consciousness; every three seconds, a child is born. This answer however is not very satisfying: most of us feel it is more nature's doing than our doing. In vitro fertilization (IVF) comes closer already, but still it's mostly nature doing its work rather than, say, some factory we've built ourselves.

How about Frankenstein's creation of the monster? Although Mary Shelley's book was and still is purely fictional, composing a human-like body from existing, dead bodyparts and bringing it to life feels closer still to creating. Perhaps this is because more human time and effort was spent. Then again, the amount of research that has gone into IVF and bringing it to the level it is today should not be underestimated either.

Finally, a computer and software that has been built from scratch feels very much a creation of our own, even though everything in the process of its creation can be traced back to nature. The biological ideas have been mostly abandoned: most people think only of electrical circuits, binary digits, and possibly positronic brains. If it ever happens, it is likely that it will have taken us over several centuries (that is, between the first ideas of a conscious computer and its realization).

¹ http://en.wikipedia.org/wiki/History_of_artificial_intelligence

Why is this last suggestion the most satisfactory of all four? Not only will we feel we have built it from scratch, we will also have been able to form the consciousness' personality; something we had little or no control over in the first three suggestions.

What is consciousness?

There is no widely accepted definition of consciousness yet. Physicists attribute a different meaning to consciousness than psychologists do. Some philosophers claim that consciousness resists or even defies definition. Amit Goswami in his book *The Self-Aware Universe* (chapter 7, p. 105) writes that the Oxford English Dictionary gives not one but six definitions of consciousness¹:

1. Joint or mutual knowledge.
2. Internal knowledge or conviction, especially of one's own ignorance, guilt, deficiencies, and so forth.
3. The fact or state of being conscious or aware of anything.
4. The state or faculty of being conscious as a condition of concomitant of all thought, feeling and volition.
5. The totality of the impressions, thoughts, and feelings which make up a person's conscious being.
6. The state of being conscious regarded as the normal condition of healthy waking life.

He writes:

Imagine a situation in which each of these different definitions comes into play. (We shall assign each definition a subscript — 1 through 6.) A bouquet of roses is delivered to you. The delivery man, you, and the sender all share consciousness₁, regarding the gift of roses. Is it in your consciousness₂ that you know the history, associations, and connotations of roses and of their meaning as a gift to you (and in this consciousness, you may or may not appreciate the gift). Your sensory experience of roses resides in your consciousness₃, whereby you are able to smell their fragrance, see their color, and feel their thorns. It is your consciousness₄, however, that enables you to attach the meanings, consider the relationships, and make the choices connected to the gift (whether to accept or reject the roses, for example). Your consciousness₅ is what makes you the unique you, as distinct from your lover and from everyone else, who responds in a particular way to the gift of roses. It is only by virtue of your consciousness₆ that you are able to receive roses, anyway, or to experience or exhibit and of the preceding states of consciousness.

It appears consciousness is not a matter of black or white; it appears to be a gradual thing. We are inclined to call a human 'more conscious' than a turtle, for example, or an average human being more conscious than an autistic.

Despite the many definitions, there are a few properties which most of us agree are common among conscious beings:

- A conscious being is able to have thoughts about itself, an "I", if only at a very low level. It should realize such things as "if I have pain, it's mine", and "I shouldn't eat myself when I'm hungry".²
- A conscious being recognizes other conscious beings. We can point at a dog and say, "that dog is conscious".
- Douglas Hofstadter writes that a conscious system has some internal representation of the outside world. A reflection of sorts, that updates itself when the outside world changes. See the section "Emergent complexity" for a more elaborate description of this.
- Not only should a conscious being have some internal representation of the outside world, it should also be able to communicate with it, and vice versa. The better we are able to communicate with a being, the more consciousness we attribute to it.

¹ I have not been able to verify this myself.

² I remember Daniel Dennett writing this somewhere, but I can't find it anymore.

Tests of consciousness

Besides the definitions, several tests of consciousness have also been developed. Among others, Wikipedia lists the following:

- The Turing test
While originally designed by Alan Turing as a test for intelligence, it could also be used to detect consciousness. A being or computer passes the Turing test if it can simulate human conversation undetectably. The being or computer is not required to speak; rather, it communicates via typed text.
- The mirror test
With the mirror test, devised by Gordon Gallup in the 1970s, one is interested in whether animals are able to recognize themselves in a mirror.

These tests aren't perfect, because there are animals and humans that fail to pass one or more tests, but are still believed to be conscious.

Where is the consciousness?

In the introduction of *The Mind's I* (p. 5), Daniel Dennett asks what “I” is. Am I my body? Or do I own my body?

If I *have* this body, then I guess I’m something *other than* this body. When I say “I own my body” I don’t mean “This body owns itself” — probably a meaningless claim. (...) In any case, I and my body seem both intimately connected and yet distinct. I am the controller; it is the controlled. Most of the time.

If I’m not my body, am I my brain? He argues that most of us feel that if our brain be transplanted into another body — ignoring all the technical difficulties — we would go with it, into that new body. Does that mean that we are our brains? He proposes two sentences:

I have a brain.
I am a brain.

In my copy of *The Mind's I*, a previous reader has scribbled a third sentence in the margin:

I contain a brain.

Which of these three do you feel is most appropriate? For me, the first option fits my feelings best. Then Dennett asks:

Once again, if you have a brain, could you trade it in for another? How could anyone detach you from your brain in a brain switch, if you always *go with* your brain in a body switch?

When looking at it this way, it seems that the “I” is closer to the brain than to the body. But is Dennett correct when he suggests that we always go with our brain? We sometimes read that someone’s personality changes after having been through an organ transplantation, because the organ carries character traits from the original person with it. This suggests that “I” is more than just our brain, and gives credibility to the very first suggestion in this chapter: I am my body.

In Dennett’s story *Where am I?* (yes, *The Mind's I*’s authors sometimes select works of their own), he fantasizes about separating his brain from his body. His brain is left in a vat at the Manned Spacecraft Center in Houston, while it is still able to communicate with his body through an ingenious system where each of the individual nerves is reconnected using a radio link — something the scientists in the story call *stretching* of the nerves. As an experiment, Daniel turns off the radio links, immediately after which he becomes dizzy and faints. After the scientists have turned the links back on, he starts asking himself: where am I? (Hence the title of the story.) Am I in the vat, or am I outside it looking at it? He very much likes to think he is in the vat, because he believes that the location of his brain defines his position. But it seems impossible for him: he sees, hears and feels outside the vat. Only when he closes his eyes can he faintly imagine himself lying in the container.

Here Dennett introduces the notion of *point of view*: there is some geographical point in space which is the center of one’s perception, and in this story, it’s the center of Dennett’s head, even though Dennett’s brain is no longer located in his head.

Now the notion of “I” has become even more vague: it can no longer be associated with a specific region of matter, or body tissue. Scientists have long ago concluded that there is no specific bit of the brain that is responsible for the consciousness, but what I have just described seems even worse than that.



Cartoon of the human brain.
© <http://science.howstuffworks.com/brain.htm>

If one's perception is so important, what would happen if Dennett's brainless body would be further decomposed, and individual bodyparts responsible for different kinds of perception — eyes, nose, ears — be placed far away from each other? Most of us have read of heard that we can get used to wearing glasses that turn the world upside down, if we give it a day or two. But this seems far worse. Would we be able to get used to this? Or would we go crazy and *lose consciousness*?

Just like we haven't been able to find a well-accepted definition of consciousness yet, we haven't been able to locate it exactly either. Followers of Buddhism claim looking for consciousness is futile. In *Beyond the Breath*, Marshall Glickman writes at the beginning of chapter 14 (p. 179):

The major "aha" insight that led to the Buddha's enlightenment, the key that freed him from his own suffering, was the realization that there is no self—no me, no my, no mine. He saw that we don't have an individual psyche or soul that migrates from lifetime to lifetime or rests in eternity as some ethereal representation of me-ness. "'I am' is a delusion," taught the Buddha. Once that delusion is gone, so is selfishness, grasping, attachment and unhappiness.

After our (so far) unsuccessful quest for consciousness' location, Buddha's teachings sound quite plausible.

How does this change our position regarding the question posed in this paper? Has it become harder to create consciousness, because we aren't sure what or where it is? Or easier, for the same reason — perhaps it will emerge 'automatically' if whatever is supposed to have consciousness is deep and complex enough, whatever consciousness may be exactly?

Reductionism

In the previous chapter we've been trying to break consciousness down to its building blocks, to no avail. In this chapter, we're going to do the opposite: start with building blocks, and build more and more complex stuff until it is conscious.

Reductionism is the belief that everything complicated can be explained by looking at the smaller parts making up this complicated thing. For a reductionist, consciousness emerges only from the physical structure of our brains, the many firing neurons and chemical reactions taking place at incredible rate.

The opposite of reductionism is *holism*. Holism is the belief that, as Hofstadter writes, "the whole is greater than the sum of its parts" (*The Mind's I*, p. 162).

Richard Dawkins is perhaps the most convincingly writing reductionist. In his book *The Selfish Gene* he writes his story of the origin of life and the *survival of the stable* (*The Mind's I*, p. 126):

There is no need to think of design or purpose or directedness. If a group of atoms in the presence of energy falls into a stable pattern it will tend to stay that way. The earliest form of natural selection was simply a selection of stable forms and a rejection of unstable ones. There is no mystery about this. It had to happen by definition.

He starts at what we consider the earth's earliest form: a large, rough chunk of rock, with mostly simple molecules floating in its vast oceans. Scientists have tried to simulate the conditions the earth was supposedly in, and found that under the influence of ultraviolet light or electricity and enough time, the atoms comprising these simple molecules regrouped into more complex ones, such as amino acids.

Dawkins claims that while large molecules nowadays wouldn't last long, they were able to live far longer in the past, because there weren't any bacteria yet that 'ate' them. This allowed the molecules to react much more often, and at some point, a very special molecule was formed:

We will call it the *Replicator*. It may not necessarily have been the biggest or the most complex molecule around, but it had the extraordinary property of being able to create copies of itself. This may seem a very unlikely sort of accident to happen. So it was.

But, as said before, molecules were able to live longer, and there was a huge amount of time available. And, as Dawkins writes, with enough time, anything improbable is bound to happen *sometime*.

After the forming of the Replicator—probably the most crucial step in his story—mutations and natural selection caused many different kinds of replicators to form, some more adept at replicating than others. The replicators began to build 'shields' and 'machines' because it helped them survive, and a conclusion of Dawkins' story is (*The Mind's I*, p. 131):

Was there to be any end to the gradual improvement in the techniques and artifices used by the replicators to ensure their own continuance in the world? (...) They did not die out, for they are past masters of the survival arts. But do not look for them floating loose in the sea; they gave up that cavalier freedom long ago. Now they swarm in huge colonies, safe inside gigantic lumbering robots, sealed off from the outside world, communicating with it by tortuous indirect routes, manipulating it by remote control. They are in you and me; they created us, body and mind; and their preservation is the ultimate rationale for our existence. They have come a long way, those replicators. Now they go by the name of genes, and we are their survival machines.

I was quite impressed when I first read this; I had never thought of it like that yet. I encourage you to read his book yourself; my short summary hardly does it justice.

Emergent complexity

The famous book *Gödel, Escher, Bach* written by Douglas Hofstadter ends each chapter with a conversation between Achilles and the Tortoise, sometimes accompanied by other characters. In one of these conversations, *Prelude...Ant Fugue*, the Anteater explains how an ant colony as a whole can be said to have some form of consciousness. While each individual ant is comparatively mindless, the subtle interactions between the ants form groups to accomplish a task. The individual groups as a whole interact too, forming yet greater groups. This goes on several times, level after level, until you're looking at the ant colony as a whole. The Anteater also explains what *castes* are: the "several different varieties of ants inside any colony", such as the queen, the workers and the soldiers. Furthermore, each ant has its own specialization, depending on age and cast, among other things. Take these three things combined — the levels of structure, the specializations and the caste distributions — and you get a 'conscious' ant colony. Hofstadter compares the colony to a brain: the individual neurons are reasonably simple, but all together all the neurons form a conscious brain.



An ant colony.
(CC) Zozakrai @ commons.wikimedia.org

In the conversation's Reflections, Hofstadter stresses the roles of brains and ant colonies as representational systems (The Mind's I, p. 192):

This concept of "representational system" is a crucial one in the book, and needs a somewhat precise definition. By "representational system" we mean an active, self-updating collection of structures organized to "mirror" the world as it evolves.

And ant colony, for example, enters 'panic' state when it is attacked. Not only should a representational system represent the world's current situation, it should also be able to take the future into account:

A representational system should be able to keep on going even if cut off from contact with the reality it is "reflecting" — although you now see that "reflection" is not quite a rich enough metaphor. The isolated representational structure should now continue to evolve, at least a probable way. Actually, a good representational system will sprout parallel branches for various possibilities that can be reasonably anticipated.

Another example of a representational system, says Hofstadter, is a country (p. 192): the individual people are part of different structures, and the structures themselves are grouped into larger structures, and so on. A country has a collective representation of the world, and acts accordingly.

This formation of complex patterns from simple building blocks is called *emergence*.¹

Not a reductionist per se but a strong believer in emergence is mathematician and writer Stephen Wolfram. In his book *A New Kind Of Science*, he claims to, well, have found a new kind of science. He describes how small systems of very simple rules can generate a large amount of complexity. He explains how regular science has often approached questions from the wrong direction: we are inclined to look for complex answers to questions, while we should look for very simple solutions instead, because simple solutions can be responsible for very complex behaviour. The most illustrating example of this kind of "emergent complexity" is probably a cellular automaton based on rule 110. A cellular automaton consists of a state and rules. A state is often an array of cells, where each cell holds a specific value, while a rule specifies how at

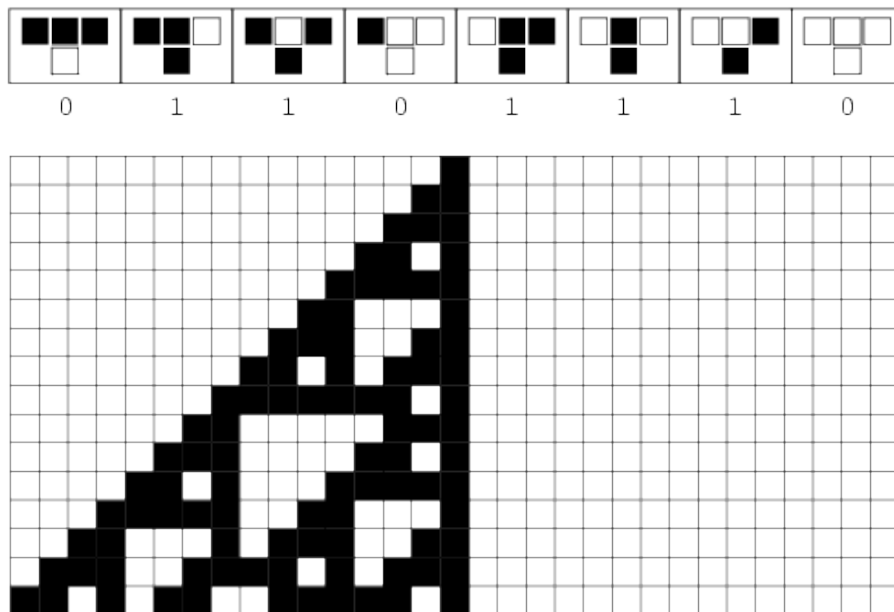
¹ See also <http://en.wikipedia.org/wiki/Emergence>

each *step* a cell's value should change, based on other cells' values, often neighbour cells. Given an initial state, the automaton evolves, and various patterns of values tend to form. Think of the Game of Life: this can be seen as an automaton with a two-dimensional state.

The family of automata of which the rule 110 automaton is a member of is one of the simplest kinds of automata imaginable: the state is a one-dimensional array of cells, and each cell has (is) either value 'black' or 'white'. At each step, each cell's new colour is a function of its old colour, and the colours of both his neighbours. This makes for $2^3 = 8$ possible situations for each cell at each step, so the automaton's rules need to define the new colour of a cell for each of these 8 possibilities.

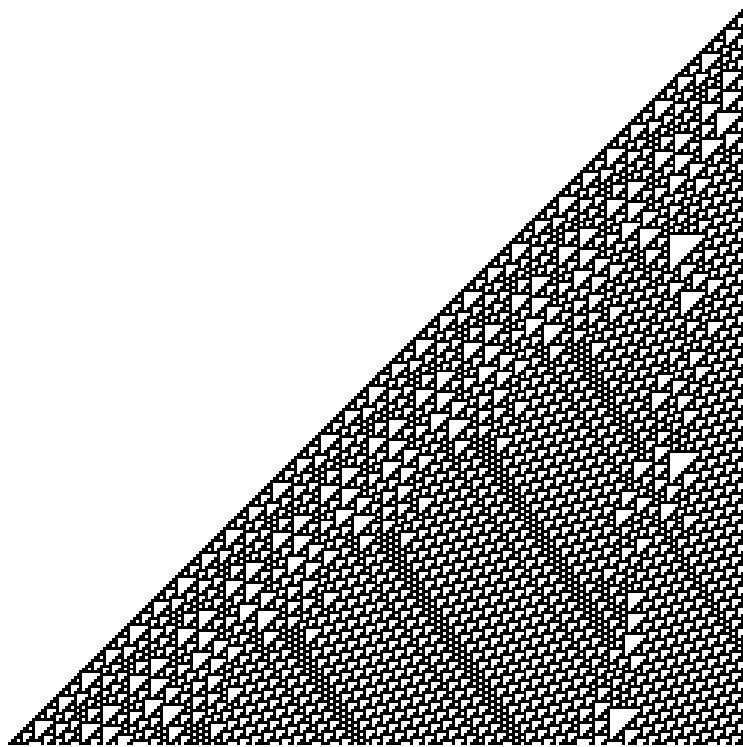
Since a cell can be only black or white, there are $2^8 = 256$ different rulesets for this specific kind of automaton. As initial condition, Wolfram uses a single black cell. Because each state is only 1-dimensional, an automaton's progress can be visualised by displaying the various arrays underneath each other. The 256 automata that are possible with the described scheme yield various results. Some 'die out' soon, leaving only white cells, or a growing line of black cells. Others yield repetitive structures, reminiscent of Pascal's triangle. And very few of the 256 yield — and this is what fascinates Wolfram so — complex, seemingly random patterns, one of which is rule 110. According to Wikipedia, Matthew Cook and Stephen Wolfram showed that the emerging structures are rich enough to support universal computation. The rest of the thousand-page *A New Kind Of Science* is devoted mainly to other systems that are simple in set-up but complex in results.

rule 110



Rule 110 and the first few states, shown underneath each other, with a single black cell as initial state (top row). The rule specifies for each three cells next to each other what value the cell under those three cells should have.

© <http://mathworld.wolfram.com/Rule110.html>



A larger part of the rule 110 cellular automaton. Note the seemingly irregular behavior.
http://en.wikipedia.org/wiki/Image:CA_rule110s.png

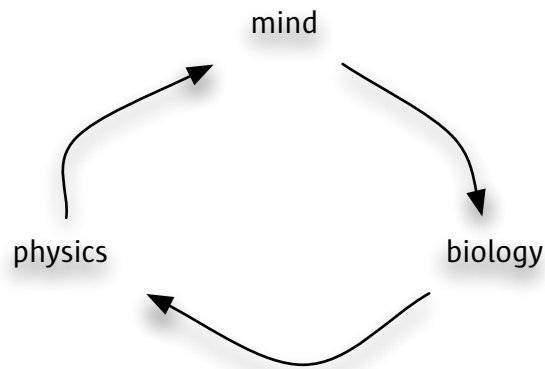
Reductionism: a vicious circle?

Over the past years, Newtonian physics have been found inadequate when trying to explain physics at the lowest level. While Newtonian physics are still taught at school and are very useful in most day-to-day situations, quantum mechanics have taken their place when science deals with particles at the atomic level. An often-used example demonstrating the nature of quantum mechanics is known as Schrödinger's cat (from Wikipedia):

A cat is placed in a sealed box. Attached to the box is an apparatus containing a radioactive atomic nucleus and a canister of poison gas. This apparatus is separated from the cat in such a way that the cat can in no way interfere with it. The experiment is set up so that there is exactly a 50% chance of the nucleus decaying in one hour. If the nucleus decays, it will emit a particle that triggers the apparatus, which opens the canister and kills the cat. If the nucleus does not decay, then the cat remains alive. According to quantum mechanics, the unobserved nucleus is described as a superposition (meaning it exists partly as each simultaneously) of "decayed nucleus" and "undecayed nucleus". However, when the box is opened the experimenter sees only a "decayed nucleus/dead cat" or an "undecayed nucleus/living cat."

In his article *Rediscovering the Mind*, Harold Morowitz describes an interesting relation between psychology, biology and physics. The reductionist view tends to argue as follows (*The Mind's I*, p. 39):

First, the human mind, including consciousness and reflective thought, can be explained by activities of the central nervous system, which, in turn, can be reduced to the biological structure and function of that physiological system. Second, biological phenomena at all levels can be totally understood in terms of atomic physics, that is, through the action and interaction of the component atoms of carbon, nitrogen, oxygen, and so forth. Third and last, atomic physics, which is now most fully understood by means of quantum mechanics, must be formulated with the mind as a primitive component of the system.



So, when trying to explain consciousness in a reductionist way, one ends up at consciousness itself! A vicious circle (or, as Amit Goswami calls it, a *tangled hierarchy*) in the reductionist philosophy? Morowitz concludes his article with:

To underrate the significance of the appearance and character of reflective thought is a high price to pay in order to honor the liberation of science from theology by our reductionist predecessors several generations back. The human psyche is part of the observed data of science. We can retain it and still be good empirical biologists and psychologists.

In the article's Reflections, Hofstadter writes:

It is an attractive notion that the mysteries of quantum physics and the mysteries of consciousness are somehow one. The epistemological loop that Morowitz describes has just about the proper amounts of hard science, beauty, weirdness, and mysticism to "sounds right." However, it is an idea that in many ways opposes an important theme of this book, which is that nonquantum-mechanical computational models of mind (and all that goes along with mind) are possible in principle.

Instead of dismissing Morowitz's idea as 'opposing our ideas', I think the loop might actually be helpful to us: if consciousness is integral to quantum physics, wouldn't it be nice if we could use it somehow when building our own consciousness? This is a very abstract idea, and while I have no idea how to realise it, I think it is worth investigating.

Objections

The Chinese Room

One of the best known argument against ‘thinking machines’ is John Searle’s Chinese Room thought experiment. Although Searle’s original description of the problem is slightly different, an intuitive way to look at the argument is this:

Suppose we have a black box that, when given texts in Chinese and questions in Chinese about that text, responds with the answers to the questions, again in Chinese. The black box contains a computer with an impressive software program on it. This is our ‘thinking machine’, because not only does the computer have to both interpret and form Chinese sentences, it also needs to be able to infer information from the story that might not be explicitly mentioned in the story.

Now replace the computer by John Searle and lock him in the black box. (Don’t worry; the box has tiny holes in it that allow fresh air to enter the box.) Searle has access to the story, the questions, and also to the program. Since Searle knows no Chinese, he understands nothing of the story and the questions; however, Searle is an able software developer and understands the program perfectly. He also has an unlimited supply of paper and ink to write notes.

Searle is now able to do the computer’s work: execute the program, maintaining the computer’s state, process the input and form the output: answers to the questions, in Chinese. Searle writes about this:

Now the claims made by strong AI are that the programmed computer understands the stories (...) it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing.

So, according to Searle, an understanding computer is impossible, more or less by definition: even if we have a robot we built from scratch that is indistinguishable from a human being on the outside (think about Andrew in Bicentennial Man), including all forms of input and output, it wouldn’t understand anything, because, well, it’s a computer.

While Searle’s argument is interesting, I don’t think it’s very useful to us: we might as well stop trying if we accepted it. Rather, we might extend the definition of “understand” to also include a ‘thinking’ machine, or instead think of a new word that means ‘to understand’ in both the human and the machine way.

Lack of nonlocality

The Self-Aware Universe, written by the American quantum physicist Amit Goswami and from which I have quoted earlier, is mainly about how at the very lowest level, everything is made of consciousness, rather than matter. In chapter 11, he explains his view on what he calls the brain-mind (p. 162):

In the past few years it has become increasingly clear to me that the only view of the brain-mind that is complete and consistent in its explanatory power is this: The brain-mind is an interactive system with both classical and quantum components.

As an example of quantum stuff going on in our minds, Goswami mentions firing neurons (p. 167):

How does an electrical impulse pass from one neuron to another across a synaptic cleft (the place where one neuron feeds another)? Conventional theory says that the synaptic transmission must be due to a chemical change. The evidence for this is somewhat circumstantial, however, and E. Harris Walker has challenged it in favor of a quantum-mechanical process. Walker thinks that the synaptic cleft is so small that the quantum

tunnelling effect may play a crucial role in the transmission of nerve signals. Quantum tunnelling is the ability that a quantum object has to pass through an otherwise impassable barrier, an ability arising from its wave nature.

Besides quantum tunnelling, another quantum effect of much importance in the book is nonlocality, defined on page 281 as “an instantaneous influence of communication without any exchange of signals through space-time; an unbroken wholeness or nonseparability that transcends space-time”. On page 21, Goswami writes:

A classical computer, notes [Richard] Feynman, can never simulate nonlocality (...) Thus, if nonlocal information processing exists in humans, it is one of our nonalgorithmic programs that a classical computer can never simulate.

Do we have nonlocal information processing? We can make a very good case for nonlocality if we accept our spirituality. Another controversial case for nonlocality is the claim of paranormal experiences. People through the centuries have claimed the capacity for telepathy, mind-to-mind transmission of information without local signals, and now there seems to be some scientific evidence for it.

Perhaps more convincing for the reductionist are the various scientific experiments Goswami describes that demonstrate nonlocality, such as the extended EEG coherence experiment (p. 171) where the brain-wave coherence of two meditative subjects is measured:

This is new evidence of quantum nonlocality. Two people meditate together; or are correlated via distant viewing, and their brain waves show coherence. Shouldn't even skeptics be intrigued?

Continuous versus discrete

Another argument against the computer's role of a mind is that minds and most of nature is continuous, while a digital computer is, well, digital, and works with discrete bits. I'm not sure if this is a strong argument. After all, a computer is made of the same kind of continuous matter as nature is, and the bits are analogue, electrical currents. Furthermore, several theories in physics describe discrete things; think of the discrete orbits electrons describe in Bohr's atom model, or the quantum jump introduced by Max Planck. It seems continuous and discrete systems are intertwined in subtle ways.

The limited brain

One of the characters in the fictional story in chapter 1 of *The Self-Aware Universe* says (p. 8):

“But you don't claim complete understanding of the brain, do you? The brain is not that simple! Didn't somebody say that if the brain were so simple that we could understand it, then we would be so simple that we couldn't?”

Is it not possible for an understanding being to completely understand itself? I'm not sure.

Other souls

So far we have searched for consciousness within ourselves. What, however, causes us to consider other beings as conscious as well?

An interesting answer to this question is described in *The Soul of Mark III Beast* by Terrel Miedaner, another selection by Dennett and Hofstadter. In it, a researcher has built a beetle-like robot called Mark III Beast which runs away when Dirksen, one of the story's characters, tries to smash it with a hammer. Earlier in the story she had remarked, "Emotionally speaking, there is a difference between the way we treat animals and the way we treat machines that defies logical explanation. I mean, I can break a machine and it doesn't really bother me, but I cannot kill an animal." (*The Mind's I*, p. 110.)

The beetle only runs away when metal comes near it, so, instructed by attorney Hunt, who has already met the beetle before, Dirksen picks it up with her bare hands and lays it on its back. Then, Hunt tells her to try and smash it again (p. 113):

Dirksen pressed her lips together tightly, raised the hammer for a final blow. But as she started to bring it down there came from the beast a sound, a soft crying wail that rose and fell like a baby whimpering. Dirksen dropped the hammer and stepped back, her eyes on the blood-red pool of lubricating fluid forming on the table beneath the creature. She looked at Hunt, horrified. "It's... it's—"

"Just a machine," Hunt said, seriously now. "Like these, its evolutionary predecessors." His gesturing hands took in the array of machinery in the workshop around them, mute and menacing watchers. "But unlike them it can sense its own doom and cry out for succor." "Turn it off," she said flatly.

The machine's 'fighting for life' together with Miedaner's use of many vivid, anthropomorphic terms throughout the story (such as "soft crying wail" in the short quotation) make the machine seem to have a soul. I believe that just like it is possible for humans to get immersed in a game or movie, it is possible for humans to start to feel for machines such as Mark III Beast. Don't we all start to feel for the little robot boy in Steven Spielberg's *A.I.*, even though we know it's a robot? Why is this? Yes, in the movie the boy fights for his life as well, and he looks and acts almost perfectly human. We like to tell ourselves, "he's not human," but this is not always so easy.

A reasonably extreme philosophical view that is related to this idea is *solipsism*: "I am the only conscious being in the universe". If I am so easily deluded by humanlike robots, what makes me so sure *you* aren't a machine too? And all the other people in the world? The only one I can be absolutely sure of is conscious is *me: cogito ergo sum*.

Goswami doesn't like being tricked by machines (*The Self-Aware Universe*, p. 23):

This reminds me of an episode of the television show "Star Trek." A con man is given an unusual punishment that on the face of it seems to be no punishment at all. He is banished to a colony where he will be the only human, surrounded by androids at his service—many in the form of beautiful maidens.

You can guess as well as I can why this is a punishment. The reason that I do not live in a solipsistic (only I am real) universe is not that others like me logically convince me of their humanness but that I have an inner connection with them. I could never have this connection with an android.

I submit that the sense we have of an inner connection with other humans is due to a real connection of the spirit. I believe that classical computers can never be conscious like us because they lack this spiritual connection.

I disagree with Goswami. I think that if a robot is lifelike enough, we will be and stay tricked by it until we for some obscure reason try to open it and find out what's on the inside. Just like a traditional artist knows there are a few specific details it should pay attention to to make a realistic drawing, I believe there are a few human traits we should simulate as good as possible in the conscious being.

When writing this, I am reminded of a few people I consider friends. They live on the other side of the earth; I only know them online. I have talked to them through instant messaging, have exchanged emails with them, have had lengthy conversations and shared a few pictures. The only information I have of them is digital information, perhaps a few megabytes of size when compressed: nothing compared to my “real life” friends. Yet I care about them nonetheless. Am I really that silly to care about *something* which is (to me) based on a bit of digital information?

Goswami might argue that I feel a long-distance, nonlocal connection with the person I’m talking to. I don’t know whether this is true. It’s worth investigating, to be sure, although it seems to me that it’s hard to falsify or proof.

Ethics

Suppose we could indeed create a conscious being. Do we have to take responsibility for such a being?

Certain people argue, “Yes! How is an artificially created conscious being different from a naturally conscious being?” A striking example is Stanislaw Lem’s story *The Seventh Sally* in which Trurl builds a complete universe in a small box and gives it to the evil king Excelsius the Tartarian to rule over (*The Mind’s I*, p. 289):

This Trurl presented to Excelsius, to rule and have dominion over forever; but first he showed him where the input and output of his brand-new kingdom were, and how to program wars, quell rebellions, exact tribute, collect taxes, and also instructed him in the critical points and transition states of that microminiaturized society—in other words the maxima and minima of palace coups and revolutions—and explained everything so well that the king, an old hand in the running of tyrannies, instantly grasped the directions and, without hesitation, while the constructor watched, issued a few trial proclamations, correctly manipulating the control knobs, which were carved with imperial eagles and regal lions.

But when Trurl returns home and tells his friend Klapaucius about his creation, Klapaucius reacts angrily:

“Enough of your boasting, not another word!” Klapaucius snapped. “Are these processes self-organizing or not?”

“Of course they are!”

“And they occur among infinitesimal clouds of electrical charge?”

“You know they do.”

“And the phenomenological events of dawns, sunsets and bloody battles are generated by the concatenation of real variables?”

“Certainly.”

“And are not we as well, if you examine us physically, mechanistically, statistically, and meticulously, nothing but the miniscule capering of electron clouds? Positive and negative charges arranged in space? And is our existence not the result of subatomic collisions and the interplay of particles, though we ourselves perceive those molecular cartwheels as fear, longing, or meditation? And when you daydream, what transpires within your brain but the binary algebra of connecting and disconnecting circuits, the continual meandering of electrons?”

“What, Klapaucius, would you equate our existence with that of an imitation kingdom locked up in some glass box?!” cried Trurl. “No, really, that’s going too far! My purpose was simply to fashion a simulator of statehood, a model cybernetically perfect, nothing more!”

“Trurl! (...) Don’t you see, when the imitator is perfect, so must be the imitation, and the semblance becomes the truth, the pretense a reality! Trurl, you took an untold number of creatures capable of suffering and abandoned them forever to the rule of a wicked tyrant... Trurl, you have committed a terrible crime!”

Although this story is pure fiction and Lem’s style of writing is somewhat amusing, the snippet makes clear that the distinction between reality and simulation is not a matter of black or white, but rather a vague, gradual thing—and also that we have to take responsibility for what we own.

What if the consciousness we have created is given enough intelligence and time to create another form of consciousness? The idea of us creating a conscious being is still long ways off, and artificial consciousness creating consciousness even more, but it is an entertaining fantasy—a weird kind of self-reference—nonetheless.

Conclusion

The biggest problem we've run into is that we're not exactly sure what we're looking for; nobody knows exactly what consciousness is, or when and how it emerges.

That said, we still have a few ideas about what a conscious being is like. After all, most people agree we all are conscious beings ourselves.

The two main views on consciousness are reductionism and holism. Reductionism says that we don't need any mysticism to get consciousness; take enough building blocks, build enough layers, and something complex enough to have consciousness will form. The emergence that is very common in nature makes this a strong claim. Holism, on the other hand, says that consciousness is more than just interacting particles and chemical reactions. The rise of quantum mechanics, with consciousness as a central but vague pillar, strengthens holism's claim.

Personally, I agree with Hofstadter and Dennett that it is possible in principle to create a conscious being. That is, a being that most humans consider conscious because of its rich ways of interaction with us. Whether it will be *really* conscious the way I *know* I am, or just *simulate* consciousness, I don't know. But is there a difference?

In any case, it will take a very long time, just like it took humans billions of years to evolve into what they currently are: curious beings with an insatiable thirst for knowledge.

Bibliography

Dennett, D.C. and Hofstadter, D.R. 1981. *The Mind's I*. Basic books.

Goswami, A. 1993. *The Self-Aware Universe*. New York: Jeremy P. Tarcher/Putnam.

Shelley, M. 1818. *Frankenstein*.

Wolfram, S. 2002. *A New Kind Of Science*.

In *The Mind's I*

Dawkins, R. 1976. *Selfish Genes and Selfish Memes*. Excerpt from *The Selfish Gene*. Oxford: Oxford University Press.

Dennett, D.C. 1978. *Where am I?* Excerpt from *Brainstorms: Philosophical Essays*. Bradford Books.

Hofstadter, D.R. 1981. *Gödel, Escher, Bach: an Eternal Golden Braid*. Basic Books, Inc.

Morowitz, H.J. *Rediscovering the Mind*. From *Psychology Today*, August 1980.

Miedaner, T. 1977. *The Soul of the Mark III Beast*. Excerpt from *The Soul Of Anna Klane*.

Turing, A.M. *Computing Machinery and Intelligence*. Appeared in *Mind*, Vol. LIX, No. 236 (1950).

Searle, J.R. 1980. *Minds, Brains and Programs*. From *The Behavioral and Brain Sciences*, vol. 3. Cambridge University Press.

In Wikipedia

<http://en.wikipedia.org/>

All pages were accessed on Friday, July 28th, 2006.

Consciousness

Emergence

History of artificial intelligence

Rule 110 cellular automaton

Schrödinger's cat

Solipsism

Meta

This paper was written using Apple's Pages 2.0.1 on a white 13" MacBook. The picture in *Reductionism: a vicious circle?* was drawn using OmniGraffle. The font used in this paper is Sun, created by Font Fabrik in Germany. It was written in Le Thoronet, in the south of France, in the very hot Summer of 2006.