Flesh and Machines: How Robots Will Change Us
by Rodney A. Brooks

2002, Pantheon Books
Chapters 7 and 8

We Are Special

We might notice our dog and make eye contact with it, cocking our head in imitation of the way it cocks its head as it looks at us. We might engage in play with our dog, waiting for it to pounce and then making some reciprocating motion. We pay attention to our dog when it greets us at the door when we get home. We would be devastated, tearful, and emotionally wrought if we saw it seriously harmed. After it has died, we might for years afterward occasionally reflect with an inner smile on some of our interactions with it.

We do not generally engage inanimate objects in the same way. But naïve people we bring off the street to interact with Kismet do engage in some forms of these activities with Kismet. Kismet and Cog have an"aliveness" about them. A few years ago, in the early days of the Cog project. Professor Sherry Turkle visited our lab. Sherry has always been something of a critic of the claims of artificial intelligence research. In recent years she has been studying how children have sharpened their distinctions between what is alive and what is not alive as they have been confronted with toys that speak, respond to stimuli, and seem to have an ongoing internal life. After visiting our lab she wrote in her book *Life on the Screen:*

> *Cog* "noticed" me soon after I entered its room. Its head turned to follow me and I was embarrassed to note that this made me happy. I found myself competing with another visitor for its attention. At one point, I felt sure that Cog's eyes had "caught" my own. My visit left me shaken—not by anything that Cog was able to accomplish but by my own reaction to "him." For years whenever I had heard Rodney Brooks speak about his robotic "creatures," I had always been careful to mentally put quotation marks around the word. But now, with Cog, I had found the quotation marks had disappeared. Despite myself and despite my continuing skepticism about this research project, I had behaved as though in the presence of another being.

Despite not wanting to be affected in this way, Turkle found herself responding involuntarily. The students who work with Kismet report the same sort of effect. They know what Kismet is capable of, and they know exactly how Kismet does what it does, since they are the ones who have programmed it. Often they are working in the same room as Kismet, and it is looking around the room or interacting with someone else. Kismet sinks into their unconsciousness as they are buried in the details of their work. Occasionally another student working on Kismet remotely might reinitialize Kismet's visual system. When this happens, Kismet goes through a calibration procedure so that its sensors can tell it which direction its eyes are pointing. To do this it scans both eyes over as far right as they can go, then as far left, then up, then down. During this time Kismet is not expressing any emotions, nor is it engaging in any of its humanlike behaviors. The students in the lab usually find this unsettling. This lifelike object that their consciousness is somehow aware of but is suppressing in detail, suddenly becomes nonlifelike.

We would most likely be unsettled if our dog suddenly started acting in the way that Kismet does during its recalibration. That sort of action is just not something we think of coming from an animate being. Clowns and mimes make use of this effect by eliminating the natural *joie de vivre* of a human and act as a machine. Should we accord Kismet the status of a being? Or is Kismet simply a machine that occasionally happens to be switched on. We can treat Kismet as something of a thought piece, because if today's Kismet itself does not qualify as worthy of having the status of a being, one can ask three further questions. First, is it possible even in principle for a machine to have the status of a being? Second, if so, then what would it have to have beyond what Kismet currently has in order to qualify as a being? Thirdly, even if we granted a machine status as a being, whatever that might mean, what sort of status should we, could we, or would we grant it?

Levels of Beingness

Before trying to answer these questions directly for Kismet or one of its descendants, let us return to our dog and then some other animals.

We are certainly all quite comfortable granting dogs *beingness.* They are living creatures, as are mice, elephants, and cats. We might debate, however, whether they are conscious. They certainly all seem to have feelings. Our dogs display fear, excitement, and contentment. It is more debatable whether they show gratitude. Most people would be willing to extend these attributes to all mammals, although mice and rats certainly seem less emotionally complex than dogs or horses. Nevertheless, we can all see a trapped house mouse exhibiting a fearful response, quivering, and breathing, and looking around for an escape route. Their fear seems visceral. We can relate to it—it is the same sort of fear we know that we feel in desperate situations. But visceral fear is not the same as reason. It is an open and often debated question whether even chimpanzees can reason. The answer to the question of whether dogs, rats, or mice can reason tends to fall along the lines of pet owners saying yes and scientists saying no.

These animals all have some aspects of beingness, though certainly not all that humans have, even forgetting for the moment their lack of syntax and technology. With this beingness we humans grant these animals certain respect and rights. While there is debate on the role of animals, and mammals in particular, for scientific experiments, almost no one argues that cruelty to animals for cruelty's sake is not immoral. In fact, there seems to be a correlation between those who treat humans with psychopathic cruelty and childhood histories of extreme cruelty to animals.

Now consider reptiles. When you quickly approach a lizard, it acts startled and runs away. It acts as if it is afraid, but are we willing to say that the lizard is really afraid? It does not have quite the same sort of reactions that a mammal has when it is afraid. A lizard has steerable eyes, and we can get some cues from it about its direction of gaze, so we can tell when it is looking in the direction of a source of danger. We may be able to see its rib cage heaving and so detect when its breathing rate increases. But is that because it is afraid, or because it is responding in a stimulus-response-like manner to danger, gearing up for a fast output of energy in a flee response? Is it really, really afraid, or does it just share some of the characteristics of being afraid that we have come to understand in mammals? Perhaps evolution built in the stimulus-response mechanism because that enabled better survival rates. Perhaps only later with mammals did evolution add *real* fear.

Like mammals and reptiles, fish are vertebrates and share a similar body plan with steerable eyes. Rather than lungs, fish have gills, and we certainly get cues about their respiration state when we see them gasping on the ground behind a fisherman on a jetty. If we are snorkeling underwater, we are totally aware of their startle response and their ability to accelerate much faster than we could ever imagine doing. But when they come to rest again a few seconds later, it is hard to judge that they have been through any traumatic or stressful experience. They seem as calm as when we first encountered them. Or perhaps it is that they are just as fearful as when we first encountered them.

If we now look at insects and arachnids, it becomes much less clear that they are ever afraid; in fact, sometimes they seem quite fearless. Insects like cockroaches will flee from oncoming footfalls, but there are no indications whatsoever of fear. Their eyes are not steerable, and many have no discernible head motion relative to their bodies. And they do not bother to breathe—they simply respire through small tunnels inward from their body surfaces. Spiders and other arachnids sometimes have steerable eyes, but they are too small for us to see them. Like insects, they often flee from danger, but they do not show any other evidence of fear. On the other hand, social insects will often attack creatures many times their size; their genes have traded off the survival of the individual, who will not reproduce in any case, for the survival of the society, in which many of their genes will be reproduced.

By the time we get down to worms and arrive finally at amoebas, we are probably all willing to say that they do not have emotional responses. They are not something to which we can map our own fear responses. They are so utterly different and alien that they and we are incomparable at the level of our emotions.

These different sorts of animals appear to have different levels of emotional responses. It also happens that we as a society tend to treat them with different levels of ethical care.

Chimpanzees are usually viewed as very humanlike. There is great debate about the cruelty of keeping chimpanzees locked in cages, and indeed our zoological gardens have changed their policies on this over the last twenty years. Nowadays chimpanzees, as well as gorillas and orangutans, the other great apes that lie on the same twig of the

evolutionary tree as humans, are normally only kept in large environments with indoor and outdoor spaces, and a freedom to roam within those expanded confines. The death of a single chimpanzee is treated very seriously and a private citizen owning a chimpanzee would be roundly condemned and ostracized were he to arbitrarily have that animal killed, no matter how swiftly or humanely. He or she might well face criminal prosecution in many parts of the world.

We treat chimpanzees with an almost, but not quite, human level of respect. We do not give them the rights of humans, but we give them a very large measure of the respect that we give to humans. There are very few research experiments done with chimpanzees and almost none that involve invasive surgery. We as a society have raised the bar very high in how we demand that there must be clear and immediate payoff for humankind before we are willing to subject chimpanzees to the sorts of experiments that we are willing to allow macaque monkeys to routinely participate in. But Marc Hauser at Harvard points out that we do not quite give chimpanzees the level of "selfness" that we give ourselves. Imagine a race of intelligent beings showing up at our planet Earth and studying us for thirty years. Imagine how you personally would feel about having a team of these aliens assigned to you who would follow you everywhere you went for the next thirty years, day and night. They would follow you to work and sit off to the side observing. They would come along on your honeymoon and follow you into your bedroom. They would silently and unobtrusively follow you into the bathroom every time you went there, and would take a small sample. I doubt you would be very comfortable with this. We are, however, quite comfortable with some of the most admired animal researchers doing just this to chimpanzees out in their African habitats.

As we move to dogs and cats and horses, there are strict laws against cruelty to them. It is the right of their owners to have them killed for no particular reason, however, as long as it is done "humanely" This usually means quickly, and without pain or emotional upset. It is marginally more acceptable socially to have a cat put down than a dog or a horse—there is still something of a question of "Why?" for the latter cases. In all three cases, however, we treat these animals with respect.

Mice and rats get much less respect. We can buy traps at our local hardware store for killing these animals. When they work completely as designed, they break the necks of the animals, which die quickly But it is not at all uncommon for the trap to break the spine of a mouse and for it to live on for some hours in pain. In our modern society that is certainly no one's business but that of the homeowner who is eliminating the pests. If, however, it became known that some particular homeowner was catching mice alive and then microwaving them to death, there would be outrage and criminal charges brought.

The level of respect we give to reptiles varies markedly from society to society By the time we get to fish there is very little argument against letting a fish suffocate to death once it has been pulled from the sea or a stream. Some societies are happy to cook fish alive, while others are more squeamish about this practice.

Many of us spray and swat insects to death without a second thought. We casually crush them with our shoes. We hang up high-temperature lamps that attract them and fry them to death. We inadvertently crush mites that live in our eyebrows and on our skin. We do not give these animals a first thought, let alone a second.

We grant animals some level of similarity to ourselves at an emotional level. That level of similarity happens to correspond fairly well to how similar they are to us in evolutionary and physiological ways.

Somewhere in that mixture of emotion and physiology we see enough similarity and have enough empathy that we treat animals that are similar to us in the moral ways that we have decided to treat other people.

Like Us

Is there, or will there ever be, enough similarity to us in our humanoid robots that we will decide to treat them in the same moral ways we treat other people and, in varying degrees, animals?

Our new humanoid robots have similar external form to us humans. That is, after all, why we call them humanoid. They have human shape and size and are able to move and respond in humanlike ways. To date, there are no robots, humanoid or otherwise, that have anything like a human physiology.

Our robots are all steel and silicon. Broadly, at least. They consist of many components made from plastics and metals. Some of them have an external skin, but they are not as soft and squishy as people. The actuators to date are

all electric motors, but the inclusion of springs and advanced control mechanisms have lead to very humanlike motions. Humanoid robot arms equipped this way are able to achieve subtle motions in response to external forces. While still clumsy by human standards they are starting to make progress toward the sorts of interaction with objects that a four-or-five-year-old can do routinely. This is very different from the actions that we have associated with robots from seeing industrial robots pick up and place objects from precise locations and with incredible speed.

The source of energy for our humanoid robots is electricity from a wall socket. There is no imperative for them to gather, store, steward, or expend energy sparingly. At this point in time they do not need to engage in so many of the behaviors that we humans engage in, almost unconsciously, every day, to maintain our bodies and our existence. We must eat and drink every few hours to survive. We must sleep on a daily basis to remain healthy. We must breathe every few seconds or we die in a matter of minutes. As long as our humanoid robots are freely plugged into a wall socket, they have no need to do any of these things. There are a few early experiments that involve having robots gain their energy from digesting biological matter and using the resulting gases to run fuel cells. But as long as our robots are still primarily steel and silicon creations, that form of energy will be a novelty It will always be possible to tap into an electricity source directly and bypass the physiological imperatives.

Sleep, however, is a different matter. The "reason" for sleep is still somewhat of a mystery. However, more and more recent studies show that at least part of the picture is that it is a time when short-term memories gathered during the waking period are consolidated into longer-term memory. It may turn out that there are legitimate reasons why our intelligent robots will have to shut down some of their interactive abilities in order to make such orderly compressions of recent experiences. If that is the case, then sleep may become something necessary to our humanoids in order for them to learn and adapt. On the other hand, it is plausible that we as engineers may be able to come up with something cleverer than evolution was able to do, and to generate algorithms that can do this consolidation while the robot is fully functional.

In the near term, the next ten to twenty years, say, it is a safe bet that our robots will remain very foreign in their physiology. It is likely that we will be able to make them more and more like humans in their external appearance. But we will always know that underneath their familiar exterior there is something very different from us.

So much for physiology. What about emotions? Are the emotions of robots at all like those of humans?

A number of robots that people have built, including Kismet and My Real Baby, are able to express emotions in humanlike ways. They use facial expressions, body posture, and prosody in their voices to express state of their internal emotions. Their internal emotions are a complex interplay of many subsystems. Some have drives, such as Kismet's *loneliness* drive, that can be satiated only by particular experiences in the world, in this case detecting a human face. Some have particular digital variables within some of their many concurrent programs with explicit names like *arousal*. There are many interactions between the emotional systems, the perceptual systems, and the motor systems.

For instance, in Kismet, when the loneliness level is high, the visual system rates things that have skin color as more interesting than the saturated colors of toys. As a result, there will be a higher tendency for Kismet to saccade toward things with skin color, which in its laboratory environment means that it will probably saccade to a human face if one present. Human hands are the other skin-colored objects in its environment, but they are usually smaller and thus less attractive than faces. If the loneliness level rises, it will tend to shift Kismet's emotional state away from happiness. Kismet will more likely become fearful or angry in response to small annoyances from moving objects near its face. It will display those emotions through its facial expressions and vocalizations. As it gets unhappier and unhappier, lonelier and lonelier, it will refuse to look at a toy even if that is the only thing in view.

This is similar to what happens in the human brain, where there are primitive centers of emotion such as the amygdala and other parts of the limbic system. These structures receive inputs from many parts of the brain's perceptual subsystems, and at the same time innervate both primitive motor sections of the brain and the more modern decision-making and reasoning centers of the brain. Antonio Damasio, a neuro-scientist in Iowa, has explained the role of these structures in popular books about the relationship between emotions and more advanced centers in the brain. His message, in brief, is that emotions are both primitive in the sense that we carry around the emotional systems that evolution installed in our brains long before we had warm blood, and that they play intimate roles in all of the higher-level decisions that we tend to think of as rational and emotionless.

So our robots are being built with emotional systems that model aspects of what goes on inside the heads (and hearts) of humans. But is "model" the operative word here? Are they real emotions, or are they only simulated

emotions? And even if they are only simulated emotions today, will the robots we build over the next few years come with real emotions? Will they need to be visceral emotions in the way that our dog can be viscerally afraid? What would it take for us to describe a response from a robot as visceral?

These are perhaps the deepest questions we, as humans, can ask ourselves about our technologies. The answers we choose to believe have significant impacts on our views of ourselves and our places in the universe. For if we accept that robots can have real emotions, we will be starting down the road to empathizing with them, and we will eventually promote them up the ladder of respect that we have constructed for animals. Ultimately we may have to worry about just what legal status certain robots will have in our society.

As we were building My Real Baby prototypes at iRobot Corporation, we built into them an emotional model. None of us were under any illusion that these were real emotions in the sense that a dog has emotions. The emotional system controls the behaviors in which the doll will engage and also triggers certain displays of emotion, so that the child knows what state the doll is in. We even called these displays *animations:* the series of facial expressions that the doll would go through, along with representative sounds (different in exact choice on every different occasion) that it makes. These were certainly fake emotions as far as we were all concerned.

Jim Lynch, a hard-nosed electrical engineer with a long career in building toys for other companies before ours, was the person responsible for designing the internal electronics for My Real Baby. As the prototypes got closer and closer to the version that was to be manufactured in bulk, we needed to do more and more testing. Within the company we set up a "baby-sitting" circle, including members of both the Toy Division and people who worked on other sorts of robots. For an hour at a time the prototype robot baby doll would be in the charge of a different person. They were to note what went wrong with it, if anything, and to try to record what had happened right before a failure. In this way we were able to detect bugs deep within the complex interactions of the hundreds of concurrent programs that generated the doll's overall behavior, and the complex electronics by which that software ran, sensed, acted.

One day Jim had just received a doll back from a baby-sitter. As it lay on the desk in his office, it started to ask for its bottle: "I want baba." It got more and more insistent as its hunger level went up, and soon started to cry. Jim looked for a bottle in his office but could not see one. He went out to the common area of the Toy Division and asked if anyone had a bottle. His doll needed one. As he found a bottle and rushed back to his office to feed the baby a realization came over him. This toy, that he had been working on for months, was different from all previous toys he had worked on. He could have ignored the doll when it started crying, or just switched it off. Instead, he had found himself *responding* to its emotions, and he had changed his behavior as though the doll had had real emotions. He was extraordinarily happy about what he had helped create. But here we see the crux of the question. Can a $100 doll have emotions, or can it merely trigger the responses that we have built into us evolutionarily, that over the eons have been honed to respond to emotions in other *real* creatures?

It is all very well for a robot to simulate having emotions. And it is fairly easy to accept that the people building the robots have included *models* of emotions. And it seems that some of today's robots and toys *appear* to have emotions. However, I think most people would say that our robots do not really have emotions. As in the cases of Sherry Turkle and Jim Lynch, our robots can possess us for a moment, making us act as if they have emotions and are beings, but we do not stay fooled for long. We treat our robots more like dust mites than like dogs or people. For us the emotions are not real.

Birds can fly. Airplanes can fly. Airplanes do not fly exactly as birds do, but when we look at them from the point of view of fluid mechanics, there are common underlying physical mechanisms that both utilize. We are not tempted to say that airplanes only simulate flying. We are all happy to say that they really do fly, although we are careful to remember that they are not flapping their wings or burning sugar in their muscles—they do not even have muscles. But they do fly.

Is our question about our robots having real emotions rather than just simulating having emotions the same sort of question as to whether both animals and airplanes fly? Or is there something deeper in the semantics of the word "emotion" as contrasted to "flying" that makes the former impossible for a machine to have but not the latter?

Specialness

Over the last four thousand years mankind has undergone some very large changes in worldview. Perhaps the issue of whether robots can have emotions, even in principle, let alone now, will be the topic of another change in worldview.

Human thought and intellectual discourse over the last five hundred years has been a history of changing understanding of mankind's place and role in the universe. Early on, the existence of a god, or gods, was postulated to explain the inexplicable. This mechanism has been very common across all human cultures. The details of the form of the gods have varied. There have been all-powerful single deities, and panoplies of flawed humanlike actors, and spirits that are at the edge of our understanding. But all these gods have had wills of their own, and have done things for their own reason, which is beyond our reason. We have been saved the need to explain everything in the universe. It has been beyond our place in the universe to understand or question the details. Religions have provided comfort where faith and a limit on questioning replaces the terrible unknown, the gnawing lack of understanding. Mysticism is comforting to us in the way it releases us from having to search further. It provides a satisfaction and an ease. It utilizes those early parts of the brain that are devoted to emotions and sends signals of contentment throughout our brain making us feel good.

In case you are explicitly religious and feel that your beliefs are being attacked by these remarks, you might perhaps reflect that they need only apply to all those other people throughout history who have worshiped other gods. Their beliefs and yours are not consistent, and so they must ultimately be mistaken. My remarks need only apply to why they are so committed to their incorrect beliefs.

And in case you are an atheist, let me remind you that you most probably have similar deep-seated beliefs, which while not openly religious share many of the same dogmatic aspects of religion. These beliefs might include faith in rationalism, science, bionomics, extropianism, or technofuturism. All of these also trigger the same sorts of satisfaction in that they provide a belief system that does not need to be questioned. They provide a place of contentment for our brains where we do not need to desperately search for answers, where we are not triggering the same responses that enabled us to think quickly and survive dangers on the savanna.

My point here is that all of us, myself included, have our own sets of deep-seated beliefs that we prefer to keep as beliefs. Our usual reaction to having them challenged is anger—a primitive fight response. Our emotional systems kick in and provoke us to try to eliminate the danger through the quickest means possible. We generally do not savor the feelings of internal conflict that thinking about our deepest-held beliefs brings upon us.

Over the last five hundred years, however, mankind as a species has been confronted again and again with challenges to our most deeply held beliefs. We have had to go through multistage processes at each of these challenges and find ways to reconcile new understandings of the universe with what we have already believed. These confrontations have always been emotional, and often they have been violent. They have taken tens to hundreds of years to become accepted as general truths, to become assimilated into the belief systems that we all hold and that provide us with the comfort levels that let us reach a state of equilibrium in our brain chemistry.

The natural and prescientific view that human cultures all seem to have adopted is that the Earth is the center of the universe, with the Sun, Moon, and stars revolving around it. As scientific observations were taken, this view was challenged. In the sixteenth century Copernicus suggested a heliocentric universe, with the Earth revolving around it.Later that century Tycho Brahe, with his reams of meticulous data, worked on showing that the planets circled the Sun but was not ready to give up the Earth as the center of the universe. Kepler, early in the seventeenth century, armed with Brahe's data, was able to break out of the belief in the simplicity of circles and show that the planets moved in elliptical orbits. The heliocentric universe by this time was an idea that was in the air, and discussed, but it came to a head in just a few years with Galileo. He was the first to turn a telescope to the sky, early in the seventeenth century, and before long the essential truth of the Sun as the center of our solar system become evident to him. This was not acceptable to the Catholic Church, and Galileo prudently chose the path of repentance rather than sticking with his truth.

The church could not accept that the terra firma on which mankind stood could not be the center of the universe. It had to be. For otherwise, why had God placed us somewhere peripheral? How could God, who had created men, for whatever reason, as the embodied masters of the universe, not place them centrally? Mankind was special. Specialness demanded a central place in the universe. Heaven was above. Hell below. The center of action was the

Earth. That was mankind's place in the universe and as such it must be the center.

To accept that the Sun was the center rather than the Earth demanded some acceptance of the lack of specialness of humans. As the idea swirled around for decades and more and more observations confirmed it, the idea was finally accepted and became mainstream. But it required accommodation in theologies, and it was the beginning of a dual belief system that many adopted. Religion taught about aspects of humanhood, while science gave the details of the universe. Mankind became separated from the physical world in some ways. As before, mankind was still special in nonphysical ways. Mankind was not subject to the same sorts of questioning and investigation as the physical world. Man's body was sacred and not subject to posthumous dissection.

Over time the church changed, and before long Jesuit priests had become master observational astronomers, chronicling the motions of the universe. Over the next few hundred years they were active in further reduction of the specialness of our home—Earth—within the universe. With the aid of telescopes it became clear that our sun was just one of many, and ultimately we learned that our sun was not at the center of those many, but off to one side in a galaxy of billions of suns. Later we realized that our galaxy was but one of billions of galaxies. The universe is vast, and we and everything we experience directly are just one tiny speck of a vast existence.

Mankind's place in the universe became distinctly nonspecial. But mankind's place on Earth was still very special. Man was the lord of his dominion, the dominion of Earth. Animals were very different from people. God had put them both on the Earth, but it was only humans that worshiped God, only humans with a special relationship with God. Furthermore, at least for Christians, God had created man in his own image. Whether that was the case, or whether just the opposite, that man had created God in his own image, need not concern us for now. Mankind had a special place in the universe. No longer the physical center of the universe, but still the center in God's view of the universe, for man had been created differently from all other creations

This special place was shaken mightily by the ideas of evolution that culminated in Charles Darwin's book *On the Origin of Species*. This was cataclysmic for people's view of themselves as being special. Not only were they closely related to apes, they were descended from the very same forebears as were the monkeys in the zoos. Monkeys and people were cousins. Images of this relationship were used in cartoons to ridicule Darwin—how could it possibly be said that humans and monkeys shared blood? Not only were all the "inferior" races of man closely related to educated people of the day, but even monkeys were kin. This loss of specialness was hard to bear.

The shock that it brought was too much for many, and even today, supposedly educated people in the United States cling to a belief that it is simply wrong. Despite the preponderance of evidence from the fossil record to experiments that any high school student can carry out in a well-equipped laboratory, people cling to the notion that it is some wrongheaded theory of scientists. The number of facts that need to be explained away by some alternative theory is simply too large. There is no rational explanation for what we see about us other than the truth of Darwin's central ideas. The details here and there may be fuzzy, but the central outline is undeniable. The last remaining bands of true disbelievers from the intellectual badlands of the United States plead for keeping our children in ignorance, as this is the only tool left to them in their battle against the tidal wave of evidence. Why do they carry out this battle? Why do they cling to their flat-Earth-like beliefs? They are afraid of losing their specialness. For those whose faith is based on literal interpretations of documents written by mystics thousands of years ago when just a few million people had ever lived, the facts discovered and accumulated over thousands of years by societies of billions of people are hard to accommodate.

Much of Western religion was able to accommodate evolution. It is hard to say whether it was the accommodators or the rejecters who were more intellectually consistent. In any case, a new model soon arose. God had created the clockwork of evolution and this was the mechanism, along with his ever present guiding hand, which had enabled the miracle that is us to evolve. There had been an *upward* movement throughout evolution toward us as the pinnacle. By the late nineteenth century there were illustrations of monkeys evolving from all fours to upright walking, through the chimpanzee and gorilla, through the "lesser races" of humans from Africa, to the top, to the special, to the white European male.

This view let people retain some specialness for themselves, but it was dead wrong when the true tenets of evolution were examined carefully. In evolution there was no special place or destination for mankind. It was all rather random, with luck playing out at every step of the way. A slightly delayed cosmic ray at some particular time might have led to an entirely different set of species on the surface of the Earth today, with mankind nowhere to be seen. This was the place for God's hand to ensure that all worked out for the best in the long run, ensuring that our special species would be here. This was how to accommodate evolution,but retain specialness.

Throughout the twentieth century there were many more blows to mankind's specialness, though none as severe as the loss of the Earth-centric universe or the loss of the direct creation of mankind by God. Perhaps in reality some of these blows should be seen as severe, but the vast majority of people have still not taken them seriously nor thought through their consequences.

To a society that was getting used to the idea that science unveiled secrets of the universe and led to an orderly explanation of all that happened, physics provided three unwelcome blows. These were the theory of relativity, quantum mechanics, and the Heisenberg uncertainty principle. Each of these in its own way weakened the idea that it was possible for a human to ultimately know what was going on in the universe. People's senses were not the ultimate arbiter of what was what in the universe. There were things that were ultimately not able to be sensed. These three realizations of physics ate away even more at the idea of the specialness of man—man could not know ultimate truths because they were unknowable.

When Crick and Watson discovered the structure of DNA, there was a further weakening of specialness. Soon it was found that all living creatures on Earth shared the same sort of DNA molecules and used the same coding scheme to transcribe sequences of base pairs into amino-acid building blocks of proteins. The proteins so produced were the mechanism of the cell that determined whether it was a bacterium, the skin of a lizard, or a human neuron in prefrontal cortex. But it was all the same language. Even worse, we humans shared regulatory genes, relatively unchanged, with animals as simple as flies. There were even clear relationships between some genes in humans and in yeast. In early 2001 the two human genome projects announced that there were probably only about 35,000 genes in a human—disappointingly few—less than twice as many as in a fruit fly. These discoveries hammer home the nonspecialness of humans. We are made of the same material as all other living things on this planet. Furthermore, our evolutionary history and relationship to all other living things is clear to see. We share some 98 percent of our genome with chimpanzees. We really are just like those animals, separated by a sliver of time in evolutionary history.

Machines

The two major blows to our specialness, that Earth is not the center of the universe and that we evolved as animals, did not come suddenly upon us. In each case, there were precursors to Galileo and Darwin, debates and vituperative arguments. These lasted for decades or centuries before the crystallizing events and for decades or centuries afterward. Indeed, it is only with the hindsight of history that we can point to Galileo's trial and Darwin's publication as the crystallizing events. They may not have seemed so out of the continuum to the participants or observers.

We have been in the beginning skirmishes of the third major challenge to our specialness for a period of almost fifty years now. We humans are being challenged by machines. Are we more than machine, or can all of our mental abilities, perceptions, intuition, emotions, and even spirituality be emulated, equaled, or even surpassed, by machines?

The machines we had been building for the last few millennia had never really challenged our specialness until recently. They were sometimes stronger or faster than us, or could fly while we could only walk and run, but animals could do all those things too.

With the invention of the computer and some clever programming of them in the 1950s and '60s, machines started to challenge us on our home turf: syntax and technology. As artificial intelligence proceeded. machines regularly started operating in arenas that had previously been the domain of unique human capabilities.

Early programs enabled computers to prove theorems in mathematics—not very hard ones at first, but proofs from the first few pages of the book *Prindpia Mathematica* by Bertrand Russell and Alfred North Whitehead, an exceedingly erudite attempt at formalizing mathematics. Computers were doing something with an intellectual flavor that only the most educated people could understand. There was also early progress on computers processing human languages, although there was a long hiccup from the mid-sixties to the mid-nineties without much outwardly visible progress. But those early programs, along with Noam Chomsky's generative grammars for all human languages, showed that perhaps language and syntax could be explained through fairly simple sets of rules, implementable on computers. By 1963 there was a computer program that could do as well as MIT undergraduates on the MIT calculus exam. By the mid-sixties, computers were playing chess well enough to compete in amateur chess tournaments.

As Hollywood caught up with what was happening in the research laboratories, there started to be movies about electronic brains, and the intelligence of machines perhaps rivaling that of humans. The questions of whether this was possible in principle, and whether it was likely technologically, gained credence. While Hollywood predicted the overwhelming of mankind by machines, a serious debate started in academia.

Those who took the strong AI position that of course machines would soon be superior to humans felt it was ridiculous that such a claim could be questioned. A number of people took up the challenge to defend the honor of humans. Many of them had good reasons to question the state of the art in computer intelligence, but they were often hoisted on their own petards by making weak counterattacks that were easy to foil.

When I was young and first read the children's book Robots *and Electronic Brains,* I noticed that although it was very enthusiastic about the technology it also had one little section showing that all was not marvelous in the world of computers. It pointed out that a few years earlier (the book was published in 1963) a "Chinese-American bookkeeper with an abacus won a race against an electronic calculating machine." Although the machines were blazingly fast, a human could still keep up! All was not lost for human intellect. As computers get roughly twice as fast every two years, any such advantage for a human, in terms of pure speed, is a losing proposition. Our computers eventually pass any fixed speed mark and are soon unimaginably faster.

Alan Turing, the British mathematician who had formalized much of computation in the thirties, wrote up a thought experiment in a 1950 paper. He started out by supposing that there was a person at the other end of a teletype (think of it as a two-person chat room with a very noisy keyboard), and you had to determine whether they were male or female by asking them questions. There was no requirement that they be truthful in answering. Then instead consider the case that you do not know whether it is a person or a computer at the other end of your link. Turing claimed that if a computer could fool people into thinking it was a person, then that would be a good test of intelligence. This has come to be known as the Turing test and has been the source of much debate, as well as a few interesting demonstrations.

Joseph Weizenbaum at MIT wrote a program in 1963 called ELIZA, which played the role of a psychiatrist. It carried out simple syntactic conversions on sentences typed into it, and did a few simple word substitutions resulting in questions that resembled those that a TV version of a Rogerian psychiatrist would ask. Although the program was simple, and clearly it did not understand language very well, some people would spend hours pouring out their hearts to it. Weizenbaum used this as evidence that the Turing test was flawed. He was distressed that people would share their intimate thoughts with the program and used its existence to argue that artificial intelligence was doomed. He had been shocked that people took his program seriously and that practicing psychiatrists, and others, speculated on the day that real psychotherapy might be provided by a machine. Worse, he was appalled at their own self-image, when they compared what might be doable by a computer to their own professional activities.

> What can the psychiatrist's image of his patient be when he sees himself, as therapist, not as an engaged human being acting as a healer, but as an information processor, following rules, etc.?

Weizenbaum was fleeing from the notion that humans were no more than machines. He could not bear the thought. So he denied its possibility outright.

Jaron Lanier, on John Brockman's www.edge.org Web site, in the year 2000 made a similarly confused claim about the futility of trying to build intelligent computers. People listen to Jaron because he is one of the founders of virtual reality and a technical wizard. He set out an argument that the Turing test is flawed because there are two ways that it can be passed. One is that computers can get smarter, or the second is that people can get dumber. He goes on to give examples of how people are slaves to machines with bad software, including those that assess our credit ratings. This, he claims, is people getting dumber to accommodate bad software—a claim that may well be valid. But then he states that this shows that "there is no epistemological difference between artificial intelligence and the acceptance of badly designed computer software."

And with that he dismisses work in artificial intelligence as being based on an intellectual mistake. Unfortunately for Jaron his conclusion does not follow from his premise. He has neither refuted the Turing test as a valid test for machine intelligence (I happen to be fairly neutral about whether it is a good test or not) nor shown that the failure of that test means that it is impossible to build intelligent machines. Weizenbaum and Lanier, both respected for their work in fields other than intelligent machines, wade into an argument and make nonsensical statements. Their aim is to discredit the idea that machines can be intelligent. My reading of them both is that they are afraid to give up on the specialness of mankind. An intelligent machine will call into question one of humankind's last bastions of

specialness, and so they deny, without rational argument, that such a machine could exist. It is intellectually too scary.

Hubert Dreyfus, a Berkeley philosopher, has fought a more explicit rearguard action against intelligent machines for over thirty-five years. In 1972 he published a book that set out what computers were incapable of doing. I think he was right about many issues, such as the way in which people operate in the world is intimately coupled to the existence of their body in that world. Unfortunately he made claims about what machines could not do in principle. He made a distinction between the nonformal aspects of human intelligence and the very formal rules that make up a computer program. Many of his arguments make such distinctions, and he explicitly stated that the nonformal aspects, the judgments, recognition of context, and subtle perceptual distinctions could not be reduced to mechanized processes. He was not able to supply any evidence for such a claim other than rightly pointing out that, at the time, AI researchers had not been able to produce that sort of informal intelligence. His mistake was to claim that these arguments were in *principle* arguments against any ultimate success. Many of the things that he said were not possible in principle have now been demonstrated.

Dreyfus was well known to the artificial intelligence community by the mid-sixties, as he had made some rather public statements about the inability of computers to play chess. He rightly pointed out that humans did not play chess the way machines were programmed to play chess in the 1960s. The machines followed the original ideas of Alan Turing, Norbert Wiener, and Claude Shannon, who had all suggested searching a tree of possible moves. Here Dreyfus's mistake was to get befuddled by the algorithmic nature of this procedure. To him it was impossible that simply following the rules of exhaustively trying moves, down to some fixed number of moves looking ahead, and evaluating the resulting positions, could generate good play. People were much more organic. They had insight. A rule-based system to play chess could not have insight.

Rather rashly Dreyfus, himself a rather poor amateur chess player claimed that no machine would ever beat him at chess. Richard Greenblatt at the MIT Artificial Intelligence Laboratory was happy to oblige him, and his MacHack program beat Dreyfus in the first and only encounter in 1967. Unbowed, Dreyfus declared that a program would never beat a good chess player, one who was nationally ranked. When that milestone was reached, Dreyfus claimed that a computer would never beat the world champion. As we all know, this did happen in 1997 when IBM's Deep Blue program beat world champion Garry Kasparov.

Dreyfus was dead right that the computer did not play chess at all like a human. But he was fooled by his own unease that intelligence could emerge from things that were described algorithmically. We saw in chapter 3 how lifelike animal behavior can arise from a complex set of interactions across a set of simple rules. Dreyfus did not understand that something could emerge from simple rules applied relentlessly that was able to do better at a task than an extremely gifted human.

We see now a pattern starting to emerge. Very bright people have drawn lines in the sand, saying that computers will not be able to achieve some sort of parity with humans. Again and again, they have had to erase those lines and draw new ones. This has happened often enough that I think a consensus has arisen that computers are able not only to calculate better than humans but to do many other tasks better than humans.

Computers are now better at doing symbolic algebra than are humans. During the sixties and seventies scientists and engineers got into the habit of using their institution's central computer to do large-scale numerical calculations. By the eighties they were even doing small-scale calculations on machines—a desktop or hand calculator. They still had to do the algebra and calculus by themselves, determining what formulas that were to be used for the calculation. That required too much insight to be done by machine. Researchers such as Jim Slagle and Joel Moses at the MIT Artificial Intelligence Laboratory worked through the sixties on programs that could manipulate symbolic mathematics, search through possibilities, and apply heuristics. Now most scientists and engineers use programs like Mathematica or MATLAB to do their symbolic algebra and calculus for them. The machines are better at it than people are, even with their insight.

Computers are also better at designing certain classes of numerical computer programs (filters for digital signal processors) than are humans. They are better at optimizing large sets of constraints, such as simultaneously setting prices for many interlocked goods and services. They are better at laying out circuits for fabrication. They are better at scheduling complex tasks or sequences of events. They are better at checking complex computer protocols for bugs. In short, they are better at many things that were previously the purview of highly trained mathematically inclined human experts. Those experts, when carrying out their work would in the past have been described as *thinking*, or *reasoning*, as they worked on their design or optimization problems.

While we have surrendered our superiority in these abilities, these formerly uniquely human abilities, to machines, we have not given up on everything. Most people would agree today that our computers and their software are not usually able to have true deep insights nor are they able to reason across very different domains, as we are apt to do.

Just as we are perfectly willing to say that an airplane can fly, most people today, including artificial intelligence researchers, are willing to say that computers, given the right set of software and the right problem domain, *can* reason about facts, *can* make decisions, and *can* have goals. People are also willing to say it is possible to build robots today that *act as if* they are afraid, or that *seem* to be afraid, or that *simulate* being afraid. However, it is much harder to find someone who will say that today's computers and robots can be *viscerally afraid.*

We draw a line at our emotions. In fact, we use derisive language about machines lacking emotions. We talk about "cold, hard machines" to indicate that they have no emotional component. Despite Kasparov's perception of Deep Blue as having made insightful plans, he was rather gleeful that despite its win it did not enjoy winning or gain any satisfaction from it.

We may have lost our central location in the universe, we may have lost our unique creation heritage different from animals, we may have been beaten out by machines in pure calculating and reason, but we still have our emotions. This is what makes us special. Machines do not have them, and we alone do. Emotions are our current last bastion of specialness. Interestingly we allow some emotions to animals—we increase our tribe to include them in the exclusion of machines from our intimate circle. Within the animal empire we are still at ease about our superior place, most of us having adapted to being related to our furry companions in our post-Darwin world.

Further Reading

Damasio, A. R. 1999. *The Feeling of What Happens: Body and Emotion in* the *Making of* Consciousness. New York: Harcourt Brace Jovanovich.

Dreyfus, H. L. 1972. *What Computers Can't Do.* New York: Harper & Row.

Dyson, G. B. 1997. *Darwin Among the* Machines. Reading, Mass.: Addison Wesley.

Greenblatt, R., D. Easdake, and S. Crocker. 1967. "The Greenblatt Chess Program." In *Proceedings of the Fall Joint Computer Conference,* pp. 801-10 Montvale, N. J.: AFIPS Press.

Turing, A. M. 1950. "Computing Machinery and Intelligence." Mind 59: 433-60. The article also appeared in G. E Luger, ed. 1995. *Computation and Intelligence: Collected Readings.* Cambridge, Mass.: MIT Press.

Turkle, S. 1995. *Life on the Screen.* New York: Simon & Schuster.

Weizenbaum, J. 1976. *Computer Power and Human Reason.* New York: Freeman.

We Are Not Special

If we accept evolution as the mechanism that gave rise to us, we understand that we are nothing more than a highly ordered collection of biomolecules. Molecular biology has made fantastic strides over the last fifty years, and its goal is to explain all the peculiarities and details of life in terms of molecular interactions. A central tenet of molecular biology is that *that is all there* is. There is an implicit rejection of mind-body dualism, and instead an implicit acceptance of the notion that mind is a product of the operation of the brain, itself made entirely of biomolecules. We will look at these statements in more detail later in the chapter. So every living thing we see around us is made up of molecules—biomolecules plus simpler molecules like water.

All the stuff in people, plants, and animals comes from transcription of DNA into proteins, which then interact with each other to produce structure and other compounds. Food, drink, and breath are also taken into the bodies of these organisms, and some of that may get directly incorporated into the organism as plaque or other detritus. The rest of it either reacts directly with the organism's biomolecules and is broken down into standard components—simple biomolecules or elements that bind to existing biomolecules—or is rejected and excreted. Thus almost everything in the body is biomolecules.

Biomolecules interact with each other according to well-defined laws. As they come together in any particular orientation, there are electrostatic forces at play that cause them to deform and physically interact. Chemical processes may be initiated that cause one or both of the molecules to cleave in interesting ways. With the totality of molecules, even in a single cell, there are chances for hundreds or thousands of different intermolecular reactions. It is impossible then to know or predict exactly which molecules will react with which other ones, but a statistical model of the likelihood of each type of reaction can be constructed. From that we can say whether a cell will grow, or whether it provides the function of a neuron, or whatever.

The body, this mass of biomolecules, is a machine that acts according to a set of specifiable rules. At a higher level the subsystems of the machine can be described in mechanical terms also. For instance, the liver takes in certain products, breaks them down, and recycles them. The details of how it operates can be divined from the particular bioreactions that go on within it, but only a few of those reactions are important to the liver itself. The vast majority of them are the normal housekeeping reactions that are in almost every cell in the body.

The body consists of components that interact according to well-defined (though not all known to us humans) rules that ultimately derive from physics and chemistry. The body is a machine, with perhaps billions of billions of parts, parts that are well ordered in the way they operate and interact. We are machines, as are our spouses, our children. and our dogs.

Needless to say, many people bristle at the use of the word "machine.' They will accept some description of themselves as collections of components that are governed by rules of interaction, and with no component beyond what can be understood with mathematics, physics, and chemistry. But that to me is the essence of what a machine is, and I have chosen to use that word to perhaps brutalize the reader a little. I want to bring home the fact that I am saying we are nothing more than the sort of machine we saw in chapter 3, where I set out simple sets of rules that can. be combined to provide the complex behavior of a walking robot. The particular material of which we are made may be different. Our physiology may be vastly different, but at heart I am saying we are much like the robot Genghis, although somewhat more complex in quantity but not in quality This is the key loss of specialness with which I claim mankind is currently faced.

And why the bristling at the word "machine"? Again, it is the deepseated desire to be special. To be more than mere. The idea that we are machines seems to make us have no free will, no spark, no life. But people seeing robots like Genghis and Kismet do not think of them as clockwork automatons. They interact in the world in ways that are remarkably similar to the ways in which animals and people interact. To an observer they certainly seem to have wills of their own.

When I was younger, I was perplexed by people who were both religious and scientists. I simply could not see how

it was possible to keep both sets of beliefs intact. They were inconsistent, and so it seemed to me that scientific objectivity demanded a rejection of religious beliefs. It was only later in life, after I had children, that I realized that I too operated in a dual nature as I went about my business in the world.

On the one hand, I believe myself and my children all to be mere machines. Automatons at large in the universe. Every person I meet is also a machine—a big bag of skin full of biomolecules interacting according to describable and knowable rules. When I look at my children, I can, when I force myself, understand them in this way I can see that they are machines interacting with the world.

But this is not how I treat them. I treat them in a very special way, and interact with them on an entirely different level. They have my unconditional love, the furthest one might be able to get from rational analysis. Like a religious scientist, I maintain two sets of inconsistent beliefs and act on each of them in different circumstances.

It is this transcendence between belief systems that I think will be what enables mankind to ultimately accept robots as emotional machines, and thereafter start to empathize with them and attribute free will, respect, and ultimately rights to them. Remarkably, to me at least my argument has turned almost full circle on itself. I am saying that we must become less rational about machines in order to get past a logical hangup that we have with admitting their similarity to ourselves. Indeed, what I am really saying is that we, all of us, overanthropomorphize humans, who are after all mere machines. When our robots improve enough, beyond their current limitations, and when we humans look at them with the same lack of prejudice that we credit humans, then too we will break our mental barrier, our need, our desire, to retain tribal specialness, differentiating ourselves from them. Such leaps of faith have been necessary to overcome racism and gender discrimination. The same sort of leap will be necessary to overcome our distrust of robots.

Resistance Is Futile

If indeed we are mere machines, then we have instances of machines that we all have empathy for, that we treat with respect, that we believe have emotions, that we believe even are conscious. That instance is us. So then the mere fact of being a machine does not disqualify an entity from having emotions. If we really are machines, then in principle we could build another machine out of matter that was identical to some existing person, and it too would have emotions and surely be conscious.

Now the question is how different can we make our Doppelganger from the original person it was modeled upon. Surely it does not have to be precisely like some existing person to be a thinking, feeling creature. Every day new humans are born that are not identical to any previous human, and yet they grow to be a unique emotional thinking, feeling creature. So it seems that we should be able to change our manufactured human a little bit and still have something we would all be willing to consider a human. Once we admit to that, we can change things some more, and some more, and perhaps eventually build something out of silicon and steel that is still functionally the same as a human, and thus would be accepted as a human. Or at least accepted as having emotions.

Some would argue that if it was made of steel and silicon, then as long as people did not know that, they might accept it as human. As soon as the secret was out, however, it would no longer be accepted. But that lack of acceptance cannot be on the basis that it is a machine, as we are already supposing that we are machines. Indeed, the many arguments that abound about why a machine can never have real emotions, or really be intelligent, all boil down to a denial of one form or another, that we are machines, or at least machines in the conventional sense,

So here is the crux of the matter. I am arguing that we are machines and we have emotions, so in principle it is possible for machines to have emotions as we have examples of them (us) that do. Being a machine does not disqualify something from having emotions. And, by straight-forward extension, does not prevent it from being conscious. This is an assault on our specialness, and many people argue that it cannot be so, and argue that they need to make the case that we are more than machines.

Sometimes they are arguing that we are more than conventional computers. This may well be the case, and I have not, so far, taken any position on this. But I *have* taken a position that we are machines.

Many hard-nosed scientists have joined this intellectual fray. Some do better than others in making their arguments. But usually those that argue that machines will never be as good as us somehow end up arguing that we are more

than mere machines. Often the players are strong materialists who proclaim that they are not arguing for any role from God, or a spirit, or a soul for people, and not even some elan vital, or life force. Rather they argue that there is something explicitly or implicitly more than a mere machine but which is at the same time of the material world. In order to make such an argument they need to invent some sort of *new stuff.* Often they deny that this is what they are doing. Below, I pick Roger Penrose, David Chalmers, and John Searle as representatives of the people who provide the most cogent arguments against people being just machines. Penrose and Chalmers could turn out to be right, although neither of them provides any data to support his hypothesis. Searle, I think, is just plain confused.

Roger Penrose, the British physicist and mathematician, is certainly a hard-nosed scientist. He has written a number of books that attack artificial intelligence research as not being likely to build intelligent, conscious machines. While he might be right about that, his arguments are flawed and are certainly good examples of a "new stuff argument.

Penrose starts out by making a mistake in understanding Godel's theorem and Turing machines. Kurt Godel shocked the world in the thirties by showing that any consistent set of axioms for mathematics had theorems that could not be proved within that set of axioms. Turing machines, the formalization of modern computers that Alan Turing came up with around the same time, operate within whatever set of axioms they are given. So combined, these two results say that there are true theorems within mathematics that a computer cannot prove to be true. So if human mathematicians were computers, there would be theorems that they could not prove. This rankles Penrose's pride in himself and his friends. They are all able to prove lots of theorems, so Penrose incorrectly concludes that humans, or at least mathematicians, cannot be computers.

Now Penrose is in a fix. He is a hard-core materialist, but the material world cannot explain what he thinks he has observed. He needs to find something new to add within the material world that is outside the realm of ordinary computers. He hits upon the microtubules inside cells where there are quantum effects in operation. He hypothesizes, with no data to support his hypothesis, that the quantum effects there are the source of consciousness. Rather than accept the idea that consciousness is just an extension of the ideas we described in chapter 3, the result of simple mindless activities coupled together, Penrose asserts that there is some inexplicable extra thing, quantum mechanics, at play in living systems.

Penrose, in his bid for scientific materialism, has resorted to a mysterious higher force. Rather than accept the offensive idea that his magnificent mind is the product of simple mechanisms playing out together, he calls into play something too complicated for us to understand fully He invents his own little deity, the god of quantum mechanics.

David Chalmers, a philosopher at the University of Santa Cruz, is another hard-nosed materialist. He invents a very different type of "new stuff." In his version there may be some fundamentally new type in the universe that we have not yet observed directly He compares it to *spin* or *charm* in particle physics—properties of subatomic particles. Neither of these can be reduced to mass or charge, or any of the more usual types we have come to understand from classical physics. In his theory there is perhaps another type like these that we cannot observe directly with our senses, just as we cannot observe spin or charm. This new type may then be the basis of consciousness. Again, we see an appeal to some sort of higher authority, something different from anything that is rule-based, or mechanismlike. Chalmers appeals to something mysterious and not understood so that he can save his materialist view of the world without having to give up his specialness, to be nothing more than a mere machine.

John Searle is a very well respected philosopher at Berkeley. He professes to believe that mind and particularly consciousness is an emergent property of our having a brain inside our skulls. He claims to be interested in scientific explanations. But deep down he reveals that he is completely unable to accept that anything but real neurons can produce consciousness and understanding. He occasionally admits as much, for example:

> In this case, we are imagining that the silicon chips have the power ... to duplicate the mental phenomena, conscious and otherwise. . . .I hasten to add that I don't for a moment think that such a thing is even remotely empirically possible. I think it is empirically absurd to suppose that we could duplicate the causal powers of neurons entirely in silicon.

Beyond this his arguments get circular, although I am sure he would rather pejoratively disagree with my analysis here. His arguments are largely of the form of imagining robots and computer programs that have the same input-output behaviors as animals or people, and then claiming that the (imagined) existence of such things proves that mental phenomena and consciousness exist as a property of the human brain, because the robot has neither of them, even though it is able to operate in the same way as a human. This argument as I have related it may not make sense to you, but I think it is a fair summary of Searle's position. And I think the real basis for his arguments, deep down,

emotionally, is that he does not want to give up the specialness of being human.

His most famous argument concerns his "Chinese Room." John Searle does not understand Chinese. He talks about the equivalent of a computer program, a set of instructions that takes inputs in Chinese and outputs answers in Chinese. Some such computer programs exist today, as long as the domain of discussion remains somewhat limited. Searle argues that if he were locked in a room with these instructions, written out in English, and a slip of paper with a question written in Chinese was passed under the door, he could follow all the rules, written up in a large book for him, and eventually pass back under the door the answer written in Chinese on a slip of paper.

Searle correctly states that he, John Searle, would still not understand Chinese. He then absurdly concludes that no computer can understand Chinese. But he is making a fundamental mistake. Just as no single neuron in a Chinese speaker understands Chinese, Searle, a component of a larger system, does not need to understand Chinese for the whole system to understand Chinese. Searle goes on to deny that the whole system, the room, and he are conscious. I claim that we just do not know whether that is possible or not. In principle, I certainly think it is possible, but perhaps there would need to be some mechanism beyond just a set of rules for Searle to interpret.

Of course, as with many thought experiments, the Chinese Room is ludicrous in practice. There would be such a large set of rules, and so many of them would need to be followed in detailed order that Searle would need to spend many tens of years slavishly following the rules and jotting down notes on an enormous supply of paper. The system—Searle and the rules—would run as a program so slowly that it could not be engaged in any normal sorts of perceptual activity. At that point it does get hard to effectively believe that the system understands Chinese by any usual understanding of "understand." But precisely because it is such a ludicrous example, slowed down by factors of billions, any conclusions from that inadequacy cannot be carried over to making conclusions about whether a computer running the same program "understands" Chinese.

My conclusion after reading Searle's arguments is that he is afraid to give a machine consciousness. He claims it is something special about human brains and neurons in particular, but he never gives a hint of what it is that makes them special. Apart from the fact that they give rise to consciousness. He never addresses why it is that a silicon-based system cannot be conscious either, as all his arguments are circular on this front.

Less sophisticated critics bash the arguments that silicon-based systems could be intelligent or conscious by pointing out that an equivalent computer could be built out of beer cans, in which each bit of information could be represented by whether a particular dedicated beer can was currently upright or upside down. Indeed such computers could be built and would be able to do precisely the same computations as any silicon computer could do, albeit thousands of millions of times slower. The argument then dismisses a set of beer cans as being intelligent as patently obvious, and so demolishes silicon computers' being intelligent. Like Searle's Chinese Room, there is no real argument made against a beer can computer being intelligent—mere ridicule is used. Sort of like the idea that the world can't possibly be round because everyone in Australia would fall off. Ridicule does not make a valid argument. Ridicule instead of reason is a well-known refuge for tribalism.

Really then, most people's arguments about whether a robot can ever be conscious or have emotions are driven by people's own emotions, by their own tribalism. As Damasio argues, our reason is driven by our emotions.. And that is true even when we are reasoning about our emotions and those of our machines.

Ultimately the arguments come down to core sets of unshakable beliefs. My own beliefs say that we are machines, and from that I conclude that there is no reason, in principle, that it is not possible to build a machine from silicon and steel that has both genuine emotions and consciousness.

Is There Something Else?

Although I have derided the use of "new stuff in attempts to explain how we are different from machines, we are still left with a dilemma, and I will now use my own "new stuff argument to hypothesize a way out of it. Even worse, I will not supply any data to back up this argument—exactly my criticism of both Penrose and Chalmers.

We know that our current robots are not as alive as real living creatures. Although Genghis can scramble over rough terrain, it does not have the long-term independence that we expect of living systems. Although Kismet can engage people in social interactions, eventually people get bored with it and some start to treat it with disdain, as though it is

an object rather than a living creature.

Are there other fields of endeavor where we are able to really make artificial systems act like real biological systems? Beyond robotics, there is another field, artificial life, or Alife, in which biology is modeled, and, as in robotics, we have made tremendous progress in building systems with interesting properties. But as with robotics, there is well-founded criticism that so far the systems are not as robust, and ultimately as lifelike, as real biological systems.

In Alife people have built systems that reproduce, inside a computer-based world. Back at the beginning of the nineties, Tom Ray, then a biologist at the University of Delaware, developed a computer program called Tierra. This system simulated a simple computer so that Ray could have complete control over how it worked. Multiple computer programs competed for resources of the processing unit in the simulated computer. Ray placed a single program in a 60,000-word memory and let it run. (His words were only five bits, to approximate the information content in three base pairs of amino acids where there are only twenty proteins coded for plus a little control.) The program tried to copy itself elsewhere in memory and then spawn off a new process to simultaneously run that program. In this way the memory soon filled with simple "creatures" that tried to reproduce by copying themselves.

Like the DNA in biological systems, the code of the computer program was used in two ways. It was interpreted to produce the process of the creature itself, and it was copied to make a child program. The simulated computer, however, was subject to two sources of error. There were "cosmic rays" that would occasionally randomly flip a bit in memory, and there were copying errors where, as a word was being written to memory, a random bit within it might flip. Thus as the memory of the simulated computer filled with copies of the original seed program, mutations appeared. Some programs simply failed to run anymore and were removed. Others started to optimize and get slightly smaller, because the scheduling policy for multiple programs implicitly had a bias for smaller programs. Before long, "parasites," less than half the size of the original seed program, evolved. They were not able to copy themselves, but they could trick a larger program into copying them rather than itself. Soon other sorts of creatures joined them in the soup, including hyper-parasites, and social programs that needed each other to reproduce.

When Tom Ray presented this work at an Alife conference in Santa Fe in 1991, there was great excitement. It seemed as though this were the key to building complex lifelike systems. Instead of having to be extremely clever, perhaps human engineers could set up a playground for artificial evolution, and interesting complex creatures would evolve. But somehow a glass ceiling was soon hit. Despite many years of further work by Ray and others, and experiments with thousands of computers connected over the Internet, nothing very much more interesting than the results of the first experiments have come along. One hypothesis about this was that the world in which the programs lived was just not complex enough for anything interesting to happen. In terms of information content, the genomes of his programs were four orders of magnitude smaller than the genome of the simplest self-sufficient cell. Also, the genotype—the coding—and the phenotype—the body—for Ray's creatures were one and the same thing. In all real biology the genotype is a strand of DNA, while the phenotype is the creature that is expressed by that set of genes.

A few years later Kari Sims came along and built an evolution system in which the genotypes and phenotypes where different. His genotypes were directed graphs with properties that allowed easy specification of symmetry and segmented body parts, such as multisegment limbs. Each element of the phenotype expressed by these genotypes was a rectangloid box that might have sensors in it, actuators to connect it to adjacent rectangloids, and a little neural network, also connected to the neural networks of adjacent body parts. A creature was formed out of many different sized rectangloids, coupled together to form legs, arms, or other body parts. The specified creature was placed in a simulated three-dimensional world with full Newtonian physics including gravity, friction, and a viscous fluid filling the volume. By changing the parameters of the fluid it could act like water, air, or a vacuum.

Sims had about a hundred creatures simulated in a generation. They would be evaluated on some metric, or fitness function, like how well they swam or crawled, and then the better ones would be allowed to reproduce. As with Ray's system, there were various ways in which mutations could happen. Over time the creatures would get better and better at whatever they were being measured for. The first sorts of creatures Sims evolved were those that could swim in water. Over time they would evolve either into long snakelike creatures or into stubbier creatures with flippers.

When these were placed on dry ground, they could not locomote very well, but by selecting for creatures that were better at locomoting, soon they evolved to be better and better at it. Some interesting results occurred along the way. Sims soon found he had to be very careful about exactly what fitness function he chose, or evolution would often

come up with something he was not expecting. An early version of the locomotion fitness function did not penalize vertical motion and was only applied to a few seconds of existence. Soon really tall creatures evolved that were good at falling over and even tumbling. They scored high marks for the amount of motion they did in a few seconds of simulated time, but they were not really locomoting in any sustainable way Then for a while creatures evolved that moved along by beating their bodies with their limbs—they had evolved to take advantage of a bug in the implementation of conservation of momentum in the simulated physics.

Eventually, Sims set the creatures a task of trying to grab a green block placed between two of them in a standardized competition format. Many different strategies quickly evolved. Some creatures blindly lashed out and tried to scoop up whatever was in the standard position with a long articulated arm. Others were more defensive, quickly deploying a shield against the opponent and then grabbing the block at their leisure Others locomoted toward the block and tried to run off into the distance, pushing the block in front of them, scooped in by a couple of stubby arms.

This work was very impressive and got people excited. It reignited the dreams inspired by Ray's work that we would be able to mindlessly evolve very intelligent creatures. Again, however, disappointment set in, as there has not been significant improvements on Sims's work in over five years.

Recently, Jordan Pollack and Hod Lipson implemented a new evolution system with similar capabilities to Sims's program. Besides evaluating their creatures in simulated Newtonian physics, they also took the extra step of connecting their creatures directly to rapid prototyping fabrication machines. Their creatures get manufactured physically in plastic, with links and ball joints. A human has to intercede to snap in electric motors in motor holders molded into the plastic. Then the creature is free of cyberspace and is able to locomote in the real world. This innovation has once again inspired people, but there is still no fundamental new idea that is letting the creatures evolve to do better and better things. The problem is that we do not really know why they do not get better and better, so it is hard to have a good idea to fix the problem.

In summary, both robots and artificial life simulations have come a long way But they have not taken off by themselves in the ways we have come to expect of biological systems. Let us assume that all the Penroses, Chalmers, and Searles are wrong. Why is it then that our models are not doing better?

There are a few hypotheses we could make about what is lacking in all our robotic and Alife models.

1. We might just be getting a few parameters wrong in all our systems.

2. We might be building all our systems in too simple environments, and once we cross a certain complexity threshold, everything will work out as we expect.

3. We might simply be lacking enough computer power.

4. We might actually be missing something in our models of biology; there might indeed be some "new stuff that we need.

The first three of these cases are all somewhat alike, although they refer to distinctly different problems. They are all alike in that, if one of them turns out to be true, it means that we do not need anyone to be particularly brilliant to get past our current problems—time and the natural process of science will see us through.

The first case, getting just a few parameters wrong, would mean that we have essentially modeled everything correctly but are just unlucky or ignorant in some minor way. If we could just stumble across the right set of parameters, or perhaps get a little deeper insight into some of the problems so that we could better choose the parameters, then things would start working better. For instance, it could be that our current neural network models will work quantitatively better if we have five layers of artificial neurons rather than today's standard three layers. Why this should be is not clear, but it is plausible that it might be so. Or it could be that artificial evolution works much better with populations of 100,000 or more, rather than the typical 1,000 or less. But perhaps these are vain hopes. One would expect that by now someone would have stumbled across a combination of parameters that worked qualitatively better than anything else around.

Now consider the second case. Although there was disappointment in the ultimate fate of the systems evolved by Ray and Sims it was the case that the environments the creatures existed in did not demand much of them. Perhaps they needed more environmental pressure in order to evolve in more interesting ways. Or perhaps we have all the

ideas and components that are needed for living, breathing robots, but we just have not put them all together all at once. Perhaps we have only operated below some important complexity threshold. While this is an attractive idea, many people have been motivated by the same line of thinking and it has only so far resulted in systems that seem to suffer from "kitchen sink" syndrome. So again, while this is possible I rank it as having low probability of being the only problem.

The third case, that of a lack of computer power, is nothing new. Artificial intelligence and artificial life researchers have perennially complained that they do not have enough computer power for their experiments. However, the amount of computer power available has continued to follow Moore's law and doubled every eighteen months or two years since the beginning of each field. So it is getting hard to justify lack computer power as the thing that is holding up progress. Of course, there has been major progress in both fields, and many areas of progress have been enabled by the gift of Moore's law. And sometimes we have indeed seen sudden qualitative changes in the way systems seem to function just because more computer power is available.

We have recently seen an example of this. After being defeated by Deep Blue, Garry Kasparov said that he was surprised by its "ability to play as though it had a plan and how it understood the essence of the position." Deep Blue was no different in essence from the earlier versions he had been playing in the late 1980s, and in fact was not very different from Richard Greenblatt's program of the mid-sixties. Deep Blue still had no strategic-planning phase, as other chess programs designed to model human playing had. It still had only a tactical search. Because of the amount of computer power it had, it had a very deep, fast tactical search. Whereas Greenblatt's program was restricted to looking at a few hundred positions per second. Deep Blue was able to look at 200 million per second. The result was a program that seemed to Kasparov to have a game plan, not because there was anything new, but because more computer power made the approach feel qualitatively different. Bob Constable at Cornell University has reported to me the same sort of qualitative change of feeling he has recently noticed as he watches his automated theorem-proving programs at work. He knows that there is nothing qualitatively new, but he says that as he observes his programs operate, the deeper searches now made by his programs give behavior that feel to him like they have much more strategic plans on how to prove theorems.

These are two good examples because in both cases we know that the details of how the programs work, to play chess and to prove theorems, are nothing like the way humans attack these same problems. There is simply not enough of the slow hardware we have in our brains to look at as many positions or propositions as these programs do. But out of that mindless set of rules comes something that appears to two non AI researchers—world leaders in their respective intellectual fields—to be reasoning and thought.

Perhaps the same can happen for all that we hold dear about our humanity Perhaps our current models of intelligence and life will become intelligent and will come to life if we can only get enough computer power. I am still doubtful that more computer power, by itself, will be sufficient, however.

I think we probably need a few Einsteins or Edisons to figure out a few things that we do not yet understand. I put my intellectual bets on the fourth case, above, that we are still missing something in biology, that there is some "new stuff."

Unlike Penrose, Chalmers, and even Searle, however, I am betting that the new stuff is something that is already staring us in the nose, and we just have not seen it yet. The essence of the argument that I will make here is that there is something that in some sense is obvious and plain to see but that we are not seeing it in any of the biological systems that we examine. Since I do not know what it is, I cannot talk about it directly Instead, I must resort to a series of analogies.

First, let us use an analogy from physics, and building physical simulators. Suppose we are trying to model elastic objects falling and colliding. If we did not quite understand physics, we might leave out mass as a specifiable attribute of objects. This would be fine for the falling behavior, since everything falls in Earth's gravity with an acceleration independent of its mass. So if we implemented that first, we might become very encouraged by how well we were doing. But then when we came to implement collisions, no matter how much we tweaked parameters or no matter how hard we computed, the system would just not work correctly. Without mass in there, the simulation will never work.

So far, this analogy is a little like Chalmers's argument. My next step will depart from Chalmers however. Chalmers's argument calls for this missing thing, mass in my example above, to be something additional to the current physics of the world. He calls for it to be disruptive of our understanding of the universe. Analogies for

what Chalmers hypothesizes in terms of how it would rock the current scientific world would be the discovery of X-rays a century ago, which led ultimately to quantum mechanics, or the discovery of the constancy of the speed of light, which lead to Einstein's theories of relativity. Both these discoveries added completely new layers to our understandings of the universe. We eventually realized that our old understanding of physics was only an approximation of what was really happening in the universe, useful at the scales at which it had been developed, but dangerously incorrect at other scales. My version of the "new stuff is not at all disruptive. It is just some new mathematics, which I have provocatively been referring to as "the juice" for the last few years. But it is not an elixir of life that I hypothesize, and it is does not require any new physics to be present in living systems. My hypothesis is that we may simply not be seeing some fundamental mathematical description of what is going on in living systems. Consequently we are leaving out the necessary generative components to produce the processes that give rise to those descriptions as we build our artificial intelligence and artificial life models.

We have seen a number of new mathematical techniques arrive with great fanfare and promise over the last thirty years. They have included *catastrophe theory, chaos theory, dynamical systems, Markov random fields,* and *wavelets,* to name a few. Each time, researchers noticed ways in which they could be used to describe what was going on inside living systems. There often seemed to be confusion about the use of these mathematical techniques. Having identified these ways of describing what was happening inside biological systems using these techniques, researchers often jumped, without comment, to using them as explanative models of how the biological systems were actually operating. They then used the techniques as the foundations of computations that were meant to simulate what was going on inside the biological systems, but without any real evidence that they were good generative models. In any case, none of these ended up providing the breakthrough improvements that the early adopter evangelists had often predicted.

None of these mathematical techniques has the right sort of systems flavor that I am hypothesizing. The closest analogy that I can come up with is that of computation. Now I am not saying that computation is the missing notion, rather that it is an analogy for the sort of notion that I am hypothesizing.

First, we can note that computation was not disruptive intellectually, although the consequences of the mathematics that Turing and von Neumann developed did have disruptive technological consequences. A late-nineteenth-century mathematician would be able to understand the idea of Turing computability and a von Neumann architecture with a few days instruction. They would then have the fundamentals of modern computation. Nothing would surprise them, or cause them to cry out in intellectual pain as quantum mechanics or relativity would if a physicist from the same era were exposed to them. Computation was a gentle, nondisruptive idea, but one that was immensely powerful. I am convinced that there is a similarly, but different, powerful idea that we still need to uncover, and that will give us great explanatory and engineering power for biological systems.

For most of the twentieth century, scientists have poked electrodes into living nervous systems and looked for correlations between signals measured and events elsewhere in the creature or its environment. Until the middle of the century this data was compared to notions of control of cybernetics, but in the second half it was compared to computation; how does the living system compute? That has been a driving metaphor in scientific research. I am reminded that, early on, the nervous system was thought to be a hydrodynamic system, and later a steam engine. When I was a child, I had a book which told me that the brain was like a telephone switching network. By the sixties, children's books were saying that the brain was like a digital computer, and then it became a massively parallel distributed computer. I have not seen one, but I would not be at all surprised to see a children's book today that said that the brain was like the World Wide Web with all its cross references and correlations. It seems unlikely that we have gotten the metaphor right yet. But we need to figure out the metaphor. In my view it is likely to be something like a mathematical formalism of something of which we can currently see all the pieces, but cannot yet understand how they fit together.

As another analogy to computation, imagine a society that had been isolated for the last hundred years and that had not invented computers. They did, however, have electricity and electrical instruments. Now suppose the scientists in this society were given a working computer. Would they be able to reduce it to a theoretical understanding of how it worked—how it kept a database, displayed and warped images on the screen, or played an audio CD, all with no notion of computation? I suspect that these isolated scientists would have to first invent the notion of computation, perhaps spurred on by correlations of signals that they saw by taking measurements on visible wires, and even inside the microprocessor chip. Once they had the right notion of computation, they would be able to make rapid progress in understanding the computer, and ultimately they would be able to build their own, even if they used some different fabrication technology. They would be able to do it because they would understand its principles of operation.

Now we have the analogies for the mathematics that I suspect we are missing. But where might we look for such mathematics? Ah, if only I knew! It is certainly the case that living systems are made up of matter, and that our current computational models of living systems do not adequately capture certain "computational" properties of that matter. Real matter cannot be created and destroyed arbitrarily. This is a constraint that is entirely missing from Alife simulations of living systems. Furthermore, all matter is doing things all the time that would be incredibly expensive to compute. Molecules are subject to forces from other molecules around them, and physics operates by continually minimizing these forces. That is how the shapes of the membranes of cells come about, how molecules migrate through solution and across barriers, and how large complex molecules fold up on themselves to take the physical shapes that are important for the way they interact with each other as the mechanisms of recognition, binding, or transcription.

But this is just one obvious place to look. The real trick will be to find the nonobvious, for if the juice hypothesis is true, that must be where it is hiding.

It might turn out that, for all the different aspects of biology that we model, there is a different juice that we are missing. For perceptual systems, say, there might be some organizing principle, some mathematical notion, that we need in order to understand how animal perception systems really work. Once we have discovered this juice, we will be able to build computer vision systems that are good at all the things they currently are not good at. These include separating objects from the background, understanding facial expressions, discriminating the living from the nonliving, and general object recognition. None of our current vision systems can do much at all in any of these areas.

Perhaps other versions of the juice will be necessary to build good explanations of other aspects of biology Perhaps different juices for evolution, cognition, consciousness or learning, will be discovered or invented and let those subfields of AI and Alife flower.

Or perhaps there will be just one mathematical notion, one juice, that will unify all these fields, revolutionize many aspects of research involving living systems, and enable rapid progress in AI and Alife.

Clever People and Aliens

Thus my version of the "new stuff argument boils down to wanting us to come up with some clever analysis of living systems and then to use that analysis in designing better machines. It is possible that there is indeed something beyond such an analysis, that there really is some "new stuff that relies on different physics than we currently understand. We have seen this happen twice in just over the last one hundred years, first with radiation, which ultimately lead to quantum mechanics, and then with relativity. Sometimes there really is new physics, and if we use that term broadly enough, it must surely cover what it is that makes things alive.

The next concern is whether people will ever be clever enough to understand the "new stuff," whether it is my weak version of just being better analysis, or whether it really does turn out to be some fundamentally new physics. What are the limits of human understanding?

Patrick Winston at MIT, in his undergraduate artificial intelligence lectures, likes to tell a story about a pet raccoon that he had when growing up in Illinois. The raccoon was very dexterous, able to manipulate mechanical fasteners and get to food that was supposed to be inaccessible. But Patrick says that it never occurred to him to wonder whether some day raccoonkind might eventually build a robot raccoon, just as capable as themselves. His parable is meant to be a caution to MIT undergraduates that perhaps they are not as smart as they like to think (and they certainly do like to think that way) and that perhaps we humans have too much hubris in our quest for artificial intelligence.

Could he be right? The downfall of Roger Penrose is that he refuses to accept limitations on the abilities of human mathematicians, and that leads him to misinterpret Godel's theorem. Perhaps there is even some superversion of Godel's theorem that says that any life-form in the universe cannot be smart enough to understand itself well enough to build a replica of itself through engineering a different technology. It is a bit of a stretch to think about the formal statement of such a theorem, but not difficult to see that in principle it might be the case that such a thing is true.

If such a thing is true, it still leaves the door open for someone smarter than us to figure out how we work, and to

build a working machine that is just as emotional and just as clever as us. In what way could someone else be intrinsically smarter than us? This again is an issue of not wanting to be not special.

Let us consider someone who is not quite as smart as all of us but does not realize it. Todd Woodward et al., psychologists at the University of Victoria in British Columbia, report on a patient, called JT, who suffered a brain hemorrhage. As with many such people, JT was left with some mental functions intact but others impaired. In JT's case he suffered from a *color agnosia*, a disruption of his ability to understand colors. His disruption, however, is most interesting for how mild it is. By comparing JT to normal control subjects Woodward and his coworkers were able to ascertain that JT had normal performance in distinguishing colors from each other. In abstract color gratings he was able to count various bars distinguished only by changes in color and not brightness. So called color-blind people are very poor at this task. JT was also able to manipulate color words, successfully verbally filling in the color names in phrases like "green with envy." So he is able to see different colors and use color words. The problem for JT comes when he must associate these two different abilities. When asked to name color patches on a screen, JT got only about half of them right, most typically getting similar colors confused—e.g., yellow and orange, or purple and pink. When asked to use color pencils to color in line drawings of familiar objects, JT was able to get about three-quarters of them right. These experiments were carried out, and some were repeated, over a period of years. Thus, JT was not improving his abilities, he was permanently damaged.

At first blush, this seems very strange indeed. Someone who can "see" colors and can use color words but cannot make the right associations between them. Surely we would all be clever enough to figure things out and simply relearn the colors? But if that is so, why can JT not? He was a functioning adult with a technical job before his hemorrhage. He was still a functioning adult person afterward, but had a deficit that the rest of us can see. How many "deficits" do the rest of us have, but that all of us in our land of the blind just do not see?

With small wiring changes in their brains, people can have strange deficits in what they can reason and learn about. As the product of evolution, it is unlikely that we are completely optimized, especially in cognitive areas. Evolution builds a hodgepodge of capabilities that are adequate for the niche in which a creature survives. It is entirely possible that with a few additional wiring changes in our "normal" brains we would have newfound capabilities. These could be newfound capabilities that we cannot currently reason about, just as with the agnosia patient. They would be capabilities that our own special reasoning, of which we humans are so proud, is not capable of reasoning about, without the wiring already in place.

As we relate to chimpanzees and macaque monkeys and their intellectual abilities, we can imagine, at least in principle, a race that has evolved with pretty much all the capabilities of our brains, but with additional wiring in place, and perhaps even with newer modules. Just as some of our modules have capabilities that are not present in chimpanzees, a *supersapiens* might have modules and capabilities that are not even latently present in us. Rather than being from Earth, perhaps supersapiens is from another planet, orbiting one of the billions of stars in one of the billions of galaxies that populate the universe. Now what will happen when supersapiens look at us? Will they see a raccoon with dexterous little hands? Will they see an impaired agnosia sufferer ("agnostic" just does not seem like the right word here) who simply cannot reason about things that are obvious? Or will they see a race of individuals capable of building artificial creatures with capabilities similar to their own?

The Question of Consciousness

Suppose that we are able to build machines in the not-too-distant future that we all agree have the emotions of a dog, say. Suppose further that the robots we build like this can act like a dog, and are as good companions for us as dogs. What then will we say about whether they are conscious or not? The question of consciousness is of great concern to irose, Chalmers, and Searie, and indeed many think that this is the key to understanding what makes us human.

This is a difficult question, for in general we cannot agree on whether any animals but us are conscious at all. While there is a gradation of emotional content, and hence empathy, which we attribute to different animals, there is no such consensus among people about the consciousness of animals.

Part of the problem is that we have no real operational definition of consciousness, apart from our own personal experience of what it is like to be us. We know that we ourselves have consciousness, and by extension we are willing to grant it to other human beings. But there is always a gnawing worry that perhaps their experience of consciousness is not the same as our own experience. Perhaps we, our own selves, really are unique, and all those

around us, despite their verbal reports, are not experiencing consciousness in the same way we do.

Once we extend this questioning to other animals, to orangutans, to dogs, to mice, to birds, to lizards, and to insects, we get progressively less sure of just how much consciousness they possess. Some people like to deny any hint of consciousness to any animals but ourselves. It becomes much more acceptable to slaughter whales for meat, or scientific research, or whatever happens to be the current rationale, if they are totally unconscious. It is also comforting to our sense of specialness if we alone are conscious.

In my opinion we are completely prescientific at this point about what consciousness is. We do not exactly know what it would be about a robot that would convince us that it had consciousness, even simulated consciousness. Perhaps we will be surprised one day when one of our robots earnestly informs us that it is conscious, and just like I take your word for your being conscious, we will have to accept its word for it. There will be no other option.

Ethical Slaves?

One of the great attractions of robots is that they can be our slaves, Mindlessly they can work for us, doing our bidding. At least this is one of the versions we see in science fiction.

But what if the robots we build have feelings? What if we start empathizing with them. Will it any longer be ethical to have them as slaves? This is exactly the conundrum that faced American slave owners. As they or their northern neighbors started to give humanhood to their slaves, it became immoral to enslave them. Once the specialness of European lineage over African lineage was erased, or at least blurred, it became unethical to treat blacks as slaves. They, but not cows or pigs, had the same right to freedom as did white people. Later a similar awakening happened concerning the status of women.

Fortunately we are not doomed to create a race of slaves that is unethical to have as slaves. Our refrigerators work twenty-four hours a day seven days a week, and we do not feel the slightest moral concern for them. We will make many robots that are equally unemotional, unconscious, and unempathetic. We will use them as slaves just as we use our dishwashers, vacuum cleaners, and automobiles today. But those that we make more intelligent, that we give emotions to, and that we empathize with, will be a problem. We had better be careful just what we build, because we might end up liking them, and then we will be morally responsible for their well-being. Sort of like children.

Further Reading

Brooks, R. A. 2001. "The Relationship Between Matter and Life." Nature 409: 409-11.

Chalmers, D. 1996. The Conscious Mind. New York: Oxford University Press.

Lipson, H., andJ. B. Pollack. 2000. "Automatic Design and Manufacture of Robotics Lifeforms," Nature 406: 974-78.

Maturana, H. R., and E J. Varela. 1987. The Tree of Knowledge: The Biological Roots of Human Understanding. New Science Library. Boston: Shambhala Publications.

Nolfi, S., and D. Floreano. 2000. Evolutionary Robotics. Cambridge, Mass.: MIT Press.

Penrose, R. 1989. The Emperor's New Mind. New York: Oxford University Press.

————. 1994. Shadows of the Mind. New York: Oxford University Press.

Ray, T. S. 1991. "An Approach to the Synthesis of Life." In Artificial Life *II.* Edited by C. G. Langton. Redwood City, Calif.: Addison-Wesley

Rosen, R. 1991. Life Itself. New York: Columbia University Press.

Searle,J. R. 1992. The Rediscovery of the Mind. Cambridge, Mass.: MIT Press.

Sims, K. 1994. "Evolving 3D Morphology and Behavior by Competition." Artificial Life 1: 353-72.

Woodward, T. S., et al. 1999. "Analysis of Errors in Color Agnosia: A Single-Case Study" Neurocose 5: 95-108.