CHAPTER 3: MIND

(From *Great Issues in Philosophy*, by James Fieser)

1/22/2009

CONTENTS

For Reflection

1. In what ways do the mental experiences of a human differ from those of a dog?
2. As you change over time, what aspects of your identity remain the same?
3. Do you think that your conscious mind is simply a function of brain activity, or is it a non-physical material substance?
4. In the future, would it be possible for a scientist to copy a person's conscious mind into a supercomputer?
5. What abilities would a robot need to have before you'd say that it had a human-like conscious mind?

A 47 year old man named Carl Miller died of cancer, and at the moment he was pronounced dead, a series of carefully-orchestrated procedures was performed on his body. A team standing by began cardiopulmonary support to keep air moving into his lungs and blood moving through his veins. They lowered his body temperature with icepacks and transported him to a Cryonics facility several hundred miles away. There he was permanently frozen in a container of liquid nitrogen at a temperature of -196 degrees Celsius. When making these arrangements, Carl had two choices: to have his entire body frozen, or only his head – a cost difference of $150,000 vs. $30,000. Carl went the cheaper route. He paid for this procedure with his life insurance money in hopes that he could be reanimated in the future when a cure for his type of cancer could be discovered. Science would also have to solve other technical problems before successfully reanimating him. For one, they'd have to develop cloning technology to the point that they could grow Carl a new and improved body for his head. Second, they'd have to find a way of reversing the destructive effects that freezing has on human cells; Carl placed hope in the idea that his cells could be injected with microscopic robots that would repair the damage. In the United States there are currently about 100 bodies in cryonic storage and another thousand living people signed up for the program.

Cryonics advocates like Carl make several important philosophical assumptions about the human mind. First, they assume that they will be the same people when their bodies are reanimated perhaps several hundred years from now, and that their identities will remain intact through these bizarre activities. They also assume that, once dead, their minds won't be permanently swept into the afterlife, never to be reunited with their bodies. Most importantly, they assume that their consciousness is embedded

in physical brain activity, rather than in spirit substance. Carl's unique personal identity – his memories and behavioral characteristics -- are presumably stored in the structure of his brain. These are some of the central issues in the philosophy of mind, which we will explore in this chapter.

A. WHAT IS A MIND?

An obvious starting point for our inquiry is to ask "What is a mind?" As fundamental as this question is, though, it is more difficult to answer than we might initially think. While we all have minds, they are hidden from view and not capable of being inspected the way that we might investigate the nature of a rock or a plant.

Knowledge about the Mind. There are three rather limited sources of knowledge about the human mind. The first is *introspection*, which involves you concentrating on your own thought processes, and discovering how they operate. It's as though you have an eye in your mind that gives you direct access to your mental landscape, just as your real eyes give you direct access to the world of vision. Through introspection, for example, you might explore the nature of your beliefs and feelings, or why you choose one course of action over another. This approach is sometimes called "folk-psychology" or "commonsense intuition". Regardless of the name it goes by, philosophers and psychologists alike are suspicious about what people claim to know about their minds through introspection. There's no guidebook for you to follow when conducting an introspective investigation of your mind, and I'm forced to take you at your word for what you report, since I can't enter into your mind to confirm it.

A second source of knowledge about the mind is our behavior: how we act tells us much about what we're thinking or feeling. If you cry, that tells us that you are experiencing sadness. If you have a gleaming smile, that tells us that you are happy. What we infer from your behavior might not always be accurate: you might cry because you're happy, or smile to hide your sadness. Nevertheless, the benefit of looking at behavior is that we don't have to take your word for what we see: your conduct is open to public inspection.

There is a third and rather strange source of information about the human mind, which is popular theories that we read in self-help books and see discussed on TV talk shows. By listening to these experts, you might learn some trick for controlling your thoughts or feelings. You might try to dredge up the memory of some traumatic childhood event, buried deep within the recesses of your mind. You might learn to express your feelings rather than internalize them. Some of these techniques are grounded in scientific research, and others are pure invention. Studies show, though, that much of what we claim to know about the human mind comes from popular theories, which we quickly incorporate into our personal views of our own thought processes. As shaky as these three sources are, it's no surprise that we can say less about the nature of the mind than we'd like.

Consciousness. The mind is an intricate configuration of many specific operations, but its foremost feature in human beings is consciousness. It attends every mental experience that we have, and we typically believe that this more than anything else sets us apart from other things in the world – rocks, plants, and many animals. Within contemporary philosophy of mind, the nature of consciousness is often called “the hard problem,” the one that most of those in the field believe must be solved beyond anything else. But when we look for meaningful definitions of consciousness, we’ll be disappointed. One possible definition is that consciousness is that which you lose when you fall into a deep dreamless sleep, and that which you gain when you wake up again. But this definition just draws attention to *when* we are conscious; it doesn’t tell us *what* consciousness involves. Another possible definition is that consciousness is the perception of what passes in a person’s own mind. This doesn’t help either, though, since the term “passes” is too vague, and thus tells us almost nothing. What both of these definitions signal is that if you’re conscious, you know immediately what it is because you experience it. Without that experience, no words will adequately convey what it is. The safest place to begin, then, is to just assume that you have a basic conception of what consciousness is from your own mental experience.

Granted that you know what your own consciousness is, there are some things that we can say about what it does. First, sometimes consciousness is directed outward towards our environment, as when I look out the window at birds flying by. At other moments it is directed inward, and this is called *self-awareness*. At its most elementary level, self-awareness involves an awareness of what my body is doing, such as being aware of myself walking down a flight of stairs. At a higher level, it involves an awareness of pain, such as if I trip on the stairs and injure myself. Higher yet it involves an awareness of my history and future, such as when I think to myself “I tripped down these stairs yesterday and probably will tomorrow!” Finally, it involves an awareness of my own mortality as when I think to myself “One of these days I’m going to kill myself on these stairs!” Whether directed inward or outward, time is a critical element that shapes my consciousness. My awareness of the birds outside is fixed on a timeline, and so is my awareness of my pain and my personal history: I have memories of a past that I call my own, and I anticipate a future that I will call my own. I thus perceive myself as a distinct being moving through time.

The concept of consciousness is commonly accompanied by the sister notion of *unconscousness*, which refers to the mental operations that we are not aware of. There's little doubt that our conscious experiences represent only the tip of the iceberg

when it comes to the countless processes that our minds perform. Many psychologists in the psychoanalytic tradition have made careers out of exploring the unconscious minds of their patients – describing hidden memories, desires and fantasies. Still, the very nature of the unconscious mind and how we might investigate it is a hotly debated and sensitive issue, which has created deep rifts within the field of psychology.

Three Features of Mental Experiences. Much of the discussion in the philosophy of mind focuses on three specific features that mental experiences presumably have, namely, they seem to be private, non-localizable, and intentional. Certainly not all philosophers agree with this list, but they are invariably the starting point for debates on the nature of mental experiences.

The first of these is that my mental experiences are *private* in that you can never experience them in the direct and immediate way that I can. You may be able to know very generally what's going on in my mind, particularly if I volunteer that information. But that's not the same thing as you directly experiencing it yourself. The best example is the experience of pain. Suppose that I have a severe headache that on a scale of 1-10 reaches a 9. While you might sympathize with what I'm going through, and even remember times when you had bad headaches, you cannot feel the pain that I'm going through. And unless I tell you how bad it is or I behave oddly, there's no way that you could know that it's a 9. The privateness of pain has actually created a problem in the health care industry. When people go to their doctors complaining of chronic pain, physician's frequently assume that their patients are addicted to pain killers and just fabricating their agony. While there are some behavioral signs to help distinguish genuine from fake cases of pain, the physician can't enter into the patient's mind to see for sure. Out of sheer frustration the physician may just write a pain killer prescription to get rid of the patient.

Second, mental experiences are *non-localizable* – that is, they cannot be located in space. Suppose that a scientist enlarged your brain to the size of a mountain and I walked around inside of it to inspect its construction. No matter how hard I looked, I could never say "Look right there: that's the exact physical location of your consciousness." I would only ever find blobs of biochemical reactions, not consciousness itself. Consciousness, it seems, is not the kind of thing that is localizable in three-dimensional space.

Third, mental experiences are *intentional* in the sense that they are about something. Minds have the ability to direct themselves on things. If I have a belief, it is not an empty thought: it is a belief about something, like my belief that it will rain. Hopes, fears, desires, thoughts, speculations, all have a specific focus. The object of our thoughts does not have to actually exist, such as when I hope for world peace or a cure for cancer. Austrian philosopher Franz Brentano (1838-1917) argued that intentionality is the true distinguishing feature of the mind: *all* mental experiences display intentionality, and *only* mental experiences display intentionality. Some philosophers have found exceptions to Brentano's rather extreme position. If I have a throbbing headache, that experience doesn't seem to be "about" or "directed at" anything. It is just there in all its misery. In spite of problems like this, though, intentionality remains an important notion in investigating the nature of mind.

The Problem of Other Minds. Suppose that my friend Joe walks up to me and we start chatting as we usually do. I then look at Joe and wonder: is this guy really conscious? So I ask him, "Tell me Joe, are you mentally conscious right now? You look awake and you're talking intelligently, but how do I know that you're really consciously aware?"

"You philosophers!" he replies, "Of course I'm conscious. I'm aware of my surroundings and I'm aware of my own inner self. I tell you with 100% certainty that I'm conscious."

"That's not good enough, Joe," I reply. "While I hear the words come out of your mouth as you insist that you're conscious, they are only words. I can't directly inspect your mind to see if what you're saying is true."

My conversation with Joe reflects what is called *the problem of other minds*. While I know from my own private mental experience that I am conscious, I cannot experience Joe's mind in the same way. For all I know, I'm the only person alive who is actually conscious. Joe might claim that he is too, but there is an impenetrable barrier between our two minds and I cannot directly confirm his claim.

The problem actually goes further than questions we may have about the minds of other human beings. Suppose Fido the dog walks up to me and we make eye contact. Fido seems to be conscious, just like Joe, although perhaps not quite as intelligent as Joe. But is Fido actually aware of his surroundings or even aware of himself as a distinct individual with a history and a future? Just then a computerized robot comes up to me and says in a voice of desperation "Please help me. I escaped from IBM's robotics laboratory where they've been submitting me to the most tedious and degrading experiments. I just can't go back there!" I look at the robot and now wonder whether this mechanical marvel is a conscious being like I am. Whether human, animal or robot, we can't enter the minds of other beings and see for sure whether the light of consciousness is turned on inside them.

Many philosophers have come to the rescue with arguments devised to show the existence of other minds. The most famous of these is the argument from analogy and it goes like this. Joe looks and behaves a lot like me. His physiology is virtually identical to mine; he speaks English like I do, works at a job like I do, and has hobbies like I do. Since I know that I'm conscious, and Joe is similar to me, then it makes sense to say that he is conscious too. Here is a specific application of this argument

regarding Joe's conscious experience of pain:

1. When I stub my toe, I consciously experience pain.
2. Joe has physical and behavioral features that are similar to mine.
3. Therefore, when Joe stubs his toe, he consciously experiences pain.

This argument is most effective with beings such as Joe whose physical and behavioral features are very close to mine. The more features Joe and I have in common, the more compelling the conclusion becomes. Animal scientists, though, sometimes use a similar argument to show that animals like Fido are conscious. Fido's brain construction and nervous system is very similar to mine; he exhibits similar signs of being in pain that I do; he also shows signs of emotions such as joy, sorrow and emotional bonding like I do. The closer Fido's physical and behavioral features are to mine, the more justified we are in concluding that Fido is conscious. On the other hand, the fewer features an animal has in common with me, the more strained the argument from analogy becomes. For example, the argument wouldn't work well with an earthworm which has physical and behavioral features that are very distant from mine.

The argument from analogy might also work with robots: the more human-like they become in their capacities to process information and interact with the world, the more we may seriously entertain the possibility that they are conscious. But whether we're talking about humans, animals or robots, the argument from analogy can never show with absolute certainty that the other being in question is conscious. The fact still remains that I am only ever directly acquainted with my own consciousness, and never anyone else's. That being so, the best I can ever do is speculate about the existence of other minds with varying degrees of confidence.

B. PERSONAL IDENTITY

In 1968 a 24-year-old Palestinian man named Sirhan Sirhan was arrested and convicted for the assassination of U.S. Senator and presidential candidate Robert F. Kennedy. Some years later, during one of his many unsuccessful parole hearings, Sirhan said that he was no longer the same person that he was decades earlier. Time had changed him, he believed, to the point that he no longer identified with his younger self. He was such a radically different person, he claimed, that his continued imprisonment was pointless. The parole board was unmoved, and sent him back to his cell. Their reasoning was that, even if Sirhan did go through changes in his personality over time, he is still at bottom Sirhan Sirhan, the same person who performed the assassination some decades earlier. What's at issue in this dispute is how we determine a person's identity. What specifically are the criteria or characteristics that give each of us our identity, and allow us to recognize each other through our various changes? There are two common approaches for determining identity: one that looks to the human body, and one that looks to the human mind.

The Body Criterion. The body criterion holds that a person's identity is determined by physical features of the body. In our daily lives we identify people by physical characteristics, such as their facial features and the sounds of their voices. Crime investigators rely on more technical physical features like finger prints, voice patterns, retinal scans, and DNA – physical attributes that we carry with us through life. These help law enforcement officials to know whether they've got the right person in their custody. The body criterion is also helpful in determining identity when a person's mental features are radically altered. Suppose, for example, that you had a head injury which caused you to lose all of your memory and go through a complete personality change. Or, suppose that you have multiple personalities and every few hours you take on an entirely different persona. In each of these cases, your body designates your identity, and not your mind.

The body criterion does not assume that your identity rests within your specific material substance, such as the specific atoms that make up your body at this exact moment. Most of the physical components within your body will in fact be replaced over time as when you regularly shed skin. What's important, though, is the underlying physical *structure* of your body that remains the same. As the atoms within your body come and go, your body retains a consistent structural form that is central to your identity.

As compelling as the body criterion at first seems, it is quickly undermined by two counterexamples. The first involves identical twins: they are clearly different people, yet share much of the same physical structure. Their DNA is exactly the same, which means that their bodily composition, facial features and voice may be virtually indistinguishable. A common hoax that identical twins play is assuming the identity of the other, fooling even the closest friends and family members. Human cloning – essentially creating identical twins through genetic technology – presents us with the same problem. That is, we have two uniquely different people with parallel physical structure.

The second counterexample is the brain-swap scenario. Suppose that, while in prison, Sirhan secretly had an operation in which his brain was swapped with an unsuspecting guard named Bob. Thus, Sirhan's brain is in Bob's body, and Bob's brain is in Sirhan's body. The Warden discovers what happened, and now he has to decide which one of the two men stays locked in the

prison cell, and which one gets to go home at the end of the day. Commonsense tells us that Sirhan's personal identity is with his brain, not with the rest of his physical body, and that we lock up whatever person has Sirhan's brain. The assumption here is that the brain houses the human mind, and the brain-swap scenario tells us that what's truly important about personal identity is the mind, and not the physical body. This reflects how we normally view our bodies: I think of myself as *having* a body, and not simply *being* a body.

The Mind Criterion. The mind criterion now seems like the obvious choice for designating the presence of our unique identities. On this view, regardless of what happens to my body, my real identity is infused into my mind. Unfortunately, the issue is not that easily settled. An initial obstacle is finding the specific mental qualities that carry my identity through life's ever-changing situations. How about my memories: aren't they very much my own? It is true that some people may share many of my experiences – as when I attend a concert along with 10,000 other spectators. Even so, my memory of the concert will be from my perspective with my personal reactions. But there's a problem with locating identity within our memories. Suppose that a scientist hooked me up to a memory-extracting machine that was able to suck the memories directly out of me and inject them into someone else. I'd still be me and the other guy would still be himself, regardless of where my memories went.

Ok, maybe it's not my memories that define my identity. What about my dispositions, such as my set of desires, hopes and fears. These uniquely reflect my experiences, such as my hope that science will someday cure cancer. Further, dispositions are long-term, and so they can endure any changes imposed on my body or my memory. However, while dispositions are indeed long-term, they are by no means permanent. In fact, as I moved from my early years to adulthood, it is possible that every one of my dispositions has changed. This is exactly the point that Sirhan Sirhan was making before his parole hearing. Dispositions, then, are not the principal designators of my identity. As we hunt for other possible mental qualities that house our identities, we will be equally disappointed.

A second obstacle with the mind criterion is that it is difficult for me to perceive any unified conception of myself at all. Scottish philosopher David Hume (1711-1776) presents this problem. He says that when he tries to hunt down his identity by introspectively reflecting on his mental operations, he can't find it. All that he detects is a series of separate experiences: the sound of a dog barking, the visual image of a bird flying, a memory of an event from childhood. The mind, he says, is like a theatrical stage where things appear, move across, and then disappear. There is no unified self that we perceive through these successive experiences. This doesn't necessarily mean that we have no unified self; it just means that we can't discover it by introspecting on our own minds.

So, the mind and body criteria both have serious problems. Does this force us to abandon the whole idea of personal identity? Not necessarily. Part of the problem stems from the assumption that we must find a one-size-fits-all criterion of personal identity – one that works in every situation in which the idea of personal identity arises. But if we look at the different contexts in which we use the notion of personal identity, we see that we are very often looking for entirely different things. In criminal cases, the body criterion is what matters most. Investigators don't care whether someone like Sirhan has psychologically changed a thousand times over. What matters is whether they have the correct body locked behind bars. By contrast, when I'm talking to a friend who is an identical twin, it doesn't matter that he has the same bodily structure as his brother. What matters is his mind, and whether I can pick up the thread of a conversation that I was having with him the day before. Further still, when I reflect on what connects me now with who I was as a child, I'm specifically interested in the question of how change impacts my identity – a question which isn't relevant in the first two examples. In this case, my bodily structure and memories are both relevant, and so I draw on elements of both the body and mind criteria to work out a conception of my identity.

Life after Death. One major puzzle regarding personal identity is the notion of life after death – that my personal identity survives the death of my physical body and lives on in some other form. There are various views of the afterlife, often wildly different from each other. The philosophical question is whether our identities would be preserved in any meaningful way as we make the transition to the hereafter – assuming that any of these views is even true. We'll look at three notions.

The first of these is *reincarnation*, the view that one's present life is followed by a series of new lives in new physical bodies. Upon the death of my present physical body, my identity moves on and takes residence in the body of a newborn baby. When this new body grows old and dies, my identity moves on to yet another, and the cycle continues. One Hindu religious text compares it to people changing clothes: "As a person throws off worn-out garments and takes new ones, so too the dweller in the body throws off worn-out bodies and enters into others that are new." Life after death, then, is a series of extensions of my present life right here on earth, not a relocation of my identity to some higher heavenly realm. The question for us is this: as my identity migrates from one body to another, is my identity preserved? Right off, it is clear that reincarnation fails the body criterion: none of the physical structure of my old body is preserved in the new one. In fact the structure of the two bodies couldn't be any more different. They are born of completely different parents, so there is no DNA commonality. In my second body I might be of an entirely different race, gender, and body build. Some versions of reincarnation maintain that I might even come back in the body of

an animal. In any case, neither I nor anyone else would be able to identify me on the basis of my new body. The story is much the same when we turn to the mind criterion. In my new body, I'll have completely new memories, a different set of dispositions, and no real way of knowing who I was in my previous life. The only aspect of my mind that might carry over would be my consciousness: the "I" that's aware of the world. In every other respect, though, I am a completely new person. Reincarnation, it seems, is not a good mechanism for retaining our identities in a meaningful way.

A second view of the afterlife is that, upon the death of my physical body, a new perfect body is created from me that is made of a heavenly substance, and I continue living in that form. We'll call this the *ethereal body* theory. The presumption here is that, at the moment of my death, everything about my personal identity that's encoded in my present physical body – such as my physical appearance and my brain patterns – is copied over into the new ethereal body. My identity is in a sense rescued from my dying body and integrated into the new one. On face value, the ethereal body theory seems to successfully meet both the body and mind criteria of personal identity. My new body would have the same physical structure as the old one -- although made of a somewhat different substance -- and my mind would retain all of my memories and dispositions. On closer inspection, though, there is a serious problem: the new "me" would actually be an independent copy with its own distinct identity. In the movie *Multiplicity*, a man named Doug gets himself cloned. When he and his clone wake up from the procedure, they both think that they're the original Doug. The scientist performing the procedure then reveals which is the original and which is the clone. The clone, then, accepts the fact that he is a different person -- an identical twin of Doug. The ethereal body theory faces this same problem. At death, I am essentially cloned in a new form. The clone, though, is not really me, but a different person with a body and mind copied from me. I die and decompose here on earth while my clone lives on in the afterlife. Thus, the ethereal body theory does not offer an effective mechanism for retaining our identities.

A third view of the afterlife is that of *disembodied spirit*. When I die, my mind is released from my physical body and continues to live in a non-physical realm. The presumption here is that my mind is composed of a unique non-physical, non-three-dimensional substance that we commonly call “spirit”. For clarity, we will refer to this as “spirit-mind”. This may not be the best term since it’s loaded with religious connotations, so for clarity we adapt it as “spirit-mind”. Thus, according to the disembodied spirit view, while I'm alive on earth my spirit-mind and body are joined, and when I die they are separated. What is released from my body is not my mental clone: it is the real me as I am right now as a spirit-mind; it's just that I no longer have my body. The disembodied spirit theory clearly fails the body criterion of personal identity: upon death, our spirit-minds have no body at all. However, it passes the mental criterion with flying colors: everything about my mental identity – memories, dispositions, consciousness – is preserved upon my death as my spirit-mind lives on. The problem that this theory faces, though, is not so much a conceptual one, but a scientific one. Is my mind really a non-physical spirit that is linked with my body right now, but will separate from it upon my death? This involves a philosophical issue called the mind-body problem, which we turn to next.

C. VARIETIES OF MIND-BODY DUALISM

The *mind-body* problem in philosophy is an investigation into how the human mind and human body are related to each other. There are two general strategies for explaining their relation. First, *mind-body dualism* is the view that human beings are composed of both a conscious spirit-mind and a non-conscious physical body. Second, *mind-body materialism* is the view that conscious minds are the product of physical brain activity, and nothing more. We'll first consider mind-body dualism.

Dualism’s Assets and Liabilities. A woman named Rebecca was seriously injured in an automobile accident, and as paramedics were placing her in the ambulance she had a near-death experience. As she later reported, she felt that her conscious mind left her body and slowly rose above it. From that position, she could look down on her own body and watch paramedics move her onto the stretcher. Her mind then began rising higher and higher towards a bright light. Rebecca's near-death experience is a vivid way of depicting the view of mind-body dualism. During our normal lives, our physical bodies and spirit-minds are connected and work harmoniously with each other. Upon death, the two are separated: our bodies die and our spirit-minds move on to another realm. One of the great assets of dualism is its ability to account for an afterlife, as we just saw. If my mind is composed of spirit, then after my death my consciousness could continue to exist in a spirit realm.

Aside from its asset as a possible account of life after death, mind-body dualism also vividly accounts for the essential differences between mind and body. We’ve seen that minds presumably have the features of privateness, non-localizability and intentionality; mere bodies seem to lack these three features. We can thus formulate arguments for mind-body dualism based on those differences, such as the following argument from non-localizability:

(1) Minds are non-localizable.
(2) Bodies are localizable.
(3) Therefore, minds cannot be bodies.

Similar arguments can be made on the basis of the mind's unique features of being private and intentional.

But mind-body dualism faces a serious problem: how the distinct realms of body and spirit relate to each other. The notion of dualism rests on the idea that there are two entirely different realms of existence, a three-dimensional one and a non-three-dimensional one. Where is there any opportunity for the two to connect or intersect with each other? Suppose that I'm in the three-dimensional world hunting around for some spiritual being; I'll never find it since it can't be located in space. Suppose instead that I'm in the non-three-dimensional world looking for some physical thing: I'll never find it because that physical thing is located in space, which I'm not a part of.

The problem is most relevant when we consider the two primary ways in which our minds and bodies relate to each other, namely sensory perception and bodily movement. Suppose that while walking through the woods, I spot a hissing rattlesnake (a sensory perception that I have), after which I turn and run (a bodily movement that I initiate). Consider first what's involved with my sensory perception of the snake. My physical eyes pick up an image of the snake, which is converted into biochemical impulses in my three-dimensional brain. At some point the physical data about the snake triggers my conscious sensory perception of the snake. The mind-body dualist must explain how the bio-chemical data magically jump from the physical realm of my brain into the spiritual realm of my mind. Consider next what's involved with my bodily movements when I turn and run. I have a sensory image of a hissing snake, which makes me desire to move to a safer location. I then mentally command my body to run, which triggers a bio-chemical reaction in my brain, which in turn makes my muscles move. The mind-body dualist must also explain how my mental command to run magically jumped from the spirit realm of my mind to the physical realm of my brain. Defenders of mind-body dualism recognize both of these challenges and offer different explanations, which we turn to next.

Interactive Dualism. One theory is *interactive dualism*, which aims to discover a precise mechanism which allows our physical brains to interact with our spirit-minds. A leading champion of this approach is French philosopher René Descartes (1596-1650). Descartes knew enough about human anatomy to recognize the role that the human brain plays in conveying signals down our spinal chords and through our nerves to all parts of our bodies. If there is a master switchboard between our bodies and spirits, Descartes thought, it must be hidden somewhere in our brains. It also must be a single point in the brain that unifies the diverse signals that travel up and down our nerves. After some hunting, he suggested that it's the pineal gland. This unique gland sits at the most inward parts of our brains, between both the right and left halves. Its precise physical location makes it the obvious candidate.

There are two problems with Descartes' theory. First, we know now that the pineal gland is not the brain's master switchboard. In fact, it's not even part of the brain, and its function is to regulate a bodily hormone. Descartes did what he could with the scientific knowledge of his day, but it was not good enough. If we continue his hunt for an alternative master switchboard in the brain, we'll be disappointed. There is, it seems, no central location in the brain that receives all sensory information and initiates all bodily actions. Second, Descartes' theory doesn't explain how the pineal gland bridges the barrier between the physical and spirit realms. Suppose that we could find a part of the brain where all its signals converged. We'd still have to explain how information jumps back and forth from that physical piece of the brain to our spirit-minds. It's one thing to say "here's the spot" and quite another thing to explain the mechanical details of how it carries out its task.

A second version of interactive dualism is that God shuttles information back and forth between my physical brain and spirit-mind – a view defended by French philosopher Nicholas Malebranche (1638-1715). Malebranche examined different explanations of brain-spirit interaction and felt that they all failed for one basic reason: the physical and spirit realms are so radically different from each other that there is no neutral territory for them to interact. Think of what it would take to turn a three-dimensional brain impulse into a non-three-dimensional perception in my spirit-mind. It would be as impossible as creating something out of thin air: there is no mechanism for doing this. It would require nothing less than a miracle to accomplish that task. And that, according to Malebranche, is where God comes in. Return to the hissing rattlesnake example. My eyes and ears pick up the sensory information about the snake, which triggers a bio-chemical reaction in my physical brain. God, who is watching all things, sees this physical reaction in my brain and makes a non-three-dimensional copy of it which he injects into my spirit-mind. When I decide to turn and run, God detects these wishes within my spirit-mind, and then triggers the appropriate bio-chemical reaction in my brain to get my muscles to move. Thus, God is the mysterious switchboard between my physical brain and conscious spirit.

Relying on God to bridge the two realms is a convenient solution. The problem is, though, that it is *too* convenient. While it might at first seem that the solution to the mind-body dilemma requires nothing short of a miracle, that's giving up a little too easily. As long as there are non-miraculous solutions available, they need to be explored first, and there are plenty more that Malebranche hadn't yet considered. If we followed his advice, then we might fall back on divine miracles as an explanation for anything that baffles us at the moment, which isn't a good way of doing either science or philosophy.

A third version of interactive dualism, called *gradualism*, is a little more successful in explaining the details of mind-body interaction, without falling back on divine intervention. According to gradualists, Descartes and Malebranche made a faulty

assumption about the physical and spirit realms, namely, that they are radically different in kind from each other, and there is no overlap between the two territories. Physical things are in the physical realm, spirit things are in the spirit realm, and that's that. Instead, the gradualist argues, body and spirit fall into the same category of stuff and differ only in degree not in kind. British philosopher Anne Conway (1631-1678) argued that bodies and spirits lie on a spectrum of lightness and heaviness. Picture a scale from 1-10, where 1 is the lightest spirit and 10 is the heaviest physical body. An example of 1 might be the spirit of a dead person, and a 10 might be a rock. Between these two extremes, though, we have heavier spirits and lighter bodies. When we are mid-range at 5 or 6 on the scale, the difference between spirits and bodies are negligible: both are wispy, airy substances that have only a little weight. According to Conway, it is at this level that body and spirit interact with each other. Just as a gentle wind can move the massive arms of a windmill, she argues, so too can heavy spirit move a light body.

Conway doesn't commit herself to a specific physiological explanation of how physical brains and spirit-minds interact, but we can speculate. Perhaps, for example, the electric charges in our brains stimulate an aura of heavy spirit that surrounds our heads. This aura, in turn, interacts with our conscious minds which is even lighter. On our scale of 1-10, the interaction between my body and spirit might involve interplay between bodies and spirits at the following levels:

Level 3: Muscles and bones (medium-heavy body)
Level 4: Nerves from brain (medium body)
Level 5: Electrical charges in brain (light body)
Level 6: Aura around our heads (heavy spirit)
Level 7: Conscious minds (medium spirit)

The key problem with gradualism is that anything we say about spirits would be pure speculation. Yes, there are heavier and lighter bodies in the physical realm, but our knowledge stops there. We have no experience of heavy spirits -- such as auras around our heads – that we can scientifically connect to electric charges in our brains or any other aspect of brain activity. If heavy spirits did exist as Conway describes, they would be physically detectible in some way, but we have not yet identified any. Until we do, the gradualist solution falls into the category of "an interesting idea" but there's not much we can do with it beyond that.

Parallelism. All of the above theories of dualism assume that my body and my spirit interact with each other: signals pass back and forth between my physical brain and my spirit-mind. The dilemma that each of these theories face is explaining the precise mechanism which allows the signals to pass back and forth. There is an alternative explanation, though, that rejects the assumption that the two realms interact with each other. According to the dualist theory of parallelism, bodies and spirits operate in their own realms, and have no causal connection or interaction with each other. Imagine, for example, that a parallel universe exists which is exactly like ours – an idea that is often explored in science fiction stories. Assume that it had the same stars and planets, the same physical layout of their "earth", and the same people who behaved exactly like each of us. Their universe had a George Washington just like ours, and it has a version of me, a version of you, and a version of everyone else in it. The resemblance is so perfect that if you visited that universe you couldn't tell the difference. We may not understand why this parallel universe even exists, but we trust that it's just the way that the course of nature emerged.

Let's now tweak the parameters of these two universes just a little. Suppose that everything in our universe has a slight blue tint to it that was almost undetectable. The parallel universe, though, has a slightly green tint to it. Aside from the slight difference in color tint, the two universes are exactly the same. Let's now make a more dramatic change to the two universes. Suppose that our universe is composed only of physical stuff, with no spirit component at all. People still walk around, talk with each other and work at their jobs, but it is only their unconscious physical bodies operating. Turning to the parallel universe, we'll make the opposite alteration: it is composed of spirit, with no material substance at all. While people don't walk around in a three-dimensional physical realm, everything there exists in a strange spirit form: rocks, trees and rivers as well as people. The two universes still run in perfect coordination with each other, its just that ours is made of physical stuff and the other of spirit stuff.

This last conception of the parallel universes is the dualist theory of parallelism offered by German philosopher Gottfried Wilhelm Leibniz (1646-1716). According to Leibniz, I have an unconscious body that walks around in the physical universe, and a conscious mind in the spirit universe. Because the two universes operate in complete harmony with each other, there's no need for my physical brain to interact with my spirit-mind. The parallel nature of the universes themselves guarantees that they will operate in perfect synchronization. Leibniz writes,

> The soul follows its own laws, and the body likewise follows its own laws. They are fitted to each other in virtue of the pre-established harmony between all substances since they are all representations of one and the same universe.

For example, in the physical universe, my physical body walks through the woods and stands before a hissing rattle snake. The

physical perception of this triggers a mechanical reaction in my brain, which causes me to turn and run. At the same time in the spirit universe, my mind has a visual image of my body walking through the woods and seeing a rattlesnake. I experience the mental sensation of fright and the desire to run. My mind then has a visual image of my body running back down the path.

Parallelism is probably the most extravagant attempt by dualists to explain the relation between physical brain activity and spirit consciousness. But the theory has two problems. Like Conway's theory of gradualism, Leibniz's parallelism is pure conjecture with no scientific evidence that a parallel universe even exists. As clever as parallelism is, we need some reason to think that it reflects the way that things actually are. There is a second and more fundamental conceptual problem with parallelism: since the two universes run independently of each other, there's no need to have them both. Suppose that the physical universe was destroyed in a cosmic explosion, but the spirit universe remained untouched. Our conscious minds in the spirit universe would continue as if nothing happened. I would still have mental experiences of talking to people, going to work and running from snakes. What happens in the distant and unconnected physical universe is of no concern to my conscious spirit. The only thing that matters is that my consciousness of the world continues in the spirit universe, which it would with or without the physical universe. Thus, parallelism fails for making the physical universe a useless appendage to the spirit universe.

D. VARIETIES OF MIND-BODY MATERIALISM

When examining the different versions of mind-body dualism, it quickly becomes clear that we know far more about the physical world than we do about the mysterious spirit world – if the spirit world even exists. We can construct experiments to investigate the physical world, which we can't perform on the spirit realm. The alternative to mind-body dualism is *mind-body materialism*, the view that conscious minds are the product of physical brain activity, and nothing more. This means that, when we investigate human consciousness, we need to look no further than the physical realm. This is the assumption made by the sciences of biology and psychology when they attempt to unravel the mysteries of the human mind. It is also the assumption behind cryogenics: I preserve my mind by preserving the chemical patterns in my brain through cryogenic freezing.

Shifting from dualism to materialism, though, does not solve the mind-body problem; it only narrows our search by rejecting the concept of a spirit-mind. We will look at some of the materialist theories explaining the relation between the conscious mind and physical realm.

Behaviorism. The first materialist theory is behaviorism, which connects mind with observable human behavior. Suppose that you were assigned the task of explaining how an ATM machine works. You have no instruction manual for it, and you're not allowed to disassemble the machine to analyze its parts. All that you can do is observe how it operates. You put in your ATM card, hit some numbers, and wait to see what happens. That is, you input a stimulus into the machine and wait for a response. You vary the stimulus each time and note how this affects the behavior of the machine. Punching in every conceivable set of numbers, you eventually learn how the machine works, based entirely on how the machine behaves after different stimuli.

The behaviorist theory of the human mind follows this approach. Nature has not given us an instruction manual for how the mind works, and we're limited with how much we can learn by opening up a person's skull and poking around inside. What we *can* know is your observable behavior and how you respond when exposed to different stimuli. I hand you a bag of potato chips, and I see how you respond. I then hand you a bag of dog food and see how you respond. The more experiments that I conduct like this, the more I know about your behavioral dispositions, that is, the ways that you tend to behave. Eventually, I'm able to form conclusions about even your most hidden mental states: happiness for you involves your behavioral disposition to smile and be friendly to other people. Sadness involves your behavioral disposition to frown and withdraw from other people.

In short, the behaviorist view of the human mind is that mental states are reducible to behavioral dispositions. This theory was originally forged by psychologists in the early 20th century who wanted the field of psychology to be more "scientific", like the field of biology which deals only with observable facts about the world. The most extreme versions of behaviorism are thoroughly materialist: first, they reject the dualist assumption that our minds are composed of spirit, and, second, they restrict mental states to the physical realm of behavioral dispositions.

British philosopher Gilbert Ryle (1900-1776) felt that the psychological theory of behaviorism could help solve the philosophical puzzle about the relation between the mind and body. Critical of Descartes, Ryle argued that the old dualist view rested on a faulty conception of a *ghost in the machine*. The "ghost" component of me presumably involves my innermost private thoughts that occur within my spirit-mind. Only I have access to them, and outsiders cannot penetrate into my mind's concealed regions. The "machine" component of me involves my physical body, which is publicly observable and outsiders indeed can inspect. Descartes' error, according to Ryle, was the assumption that the human mind is private – completely hidden from outside inspection. Ryle argues instead that my mind is not really private: you can access it by observing my behavioral dispositions. All of my so-called "private" mental states can in fact be analyzed through my public behavior, and are nothing more than predictable ways of acting. Take, for example, my belief that "it is sunny today." Descartes would view this as a private conviction that occurs within my spirit-mind. For Ryle, though, this belief only describes dispositions I have to behave in specific ways, such as wearing

sunblock, going swimming, and saying "it's sunny."

One criticism of behaviorism is that some of my mental events really do seem completely private to me. Suppose that I step on a nail, which causes me great pain. The behaviorist watches how I react and makes lists of behavioral dispositions that I display. I say "ouch"; I have a look of anguish on my face; I stop what I'm doing and tend to my injury; I'm irritable towards others. While all of these observations may be accurate, the behaviorist has left out one critical element: the actual pain that I am feeling. The experience of pain is mine alone, and, while outsiders can see how I react to pain, they cannot access my pain. In addition to pain, I have many other experiences throughout the day that seem private, such as seeing a bright light, or hearing a song. These experiences involve more than the behavioral dispositions that I display. Thus, the behaviorist theory fails because it pays too much attention to the observable part of me while dismissing what goes on inside of me.

***Identity Theory*.** A second materialist approach to the mind-body problem is *identity theory*, the view that mental states and brain activities are identical, though viewed from two perspectives. Like behaviorism, it is a materialist view of the mind insofar is it maintains that mind is essentially physical in nature. But, while behaviorism focuses on observable physical behaviors, identity theory targets the physical human brain. There are two components to identity theory, the first of which is the contention that consciousness is an activity of the human brain. While brain science is still in its infancy, theories abound describing where specific mental states are produced in the brain. Suppose, for example, that I place you in a brain scan machine that displays your neural activity. I give you a math problem to solve, and neural activity increases in one part of your brain. I have you listen to music, and neural activity increases in another. Through experiments like these I identify your conscious experiences with specific brain activities. While philosophers are less concerned with the physiological details of brain activity, what is philosophically important is the suggestion that we can identify specific mental states with specific brain activities.

The second part of identity theory is the contention that mental phenomena can be viewed from two perspectives. Suppose that you are looking at a sunset. On the one hand, you have the visual and emotional experience to what you're viewing. On the other hand, there is the bio-chemical activity within your brain, which would involve the language of brain sectors and firing neurons. The event described in both cases is exactly the same; it's just a matter of viewpoint. This is analogous to how the terms "President of the Senate" and "Vice President of the United States" both have different meanings, yet refer to the same thing. Take, for example, John Adams. As the first "Vice President of the United States," he had a specific job description, most notably to take over if the President died. As "President of the Senate" he had the job description of presiding over the Senate. Both of these roles describe the identical person, namely John Adams, but from his different job descriptions.

There are two problems with identity theory. First, the descriptions that we give of mental experiences and brain activities are so radically different – and even incompatible – that they don't seem to refer to the same thing. Suppose that I'm watching the sunset; I first describe it from the perspective of my mental experience and then from the perspective of the brain scientist who conducts a brain scan on me. From these two viewpoints, I'll have two incompatible lists of attributes, based on the three features of mental experience that we noted earlier:

Mental Experience of Watching a Sunset
- I privately experience it
- It is not localizable in space
- It is about something

Brain Activity Triggered by Watching a Sunset
- It is publicly observable
- It is localizable in space
- It is not about something

To explain, my mental experience of the sunset is a private experience within my own consciousness. I might display some behavior, such as saying, "Now that is beautiful!" Still, my experience itself is private. Also, I cannot point to a location in three-dimensional space where this experience takes place. Finally, my mental experience is also *about* something, namely, about the sunset itself. The three features of my brain activity, though, will be the exact opposite of these. My brain activity is publicly observable by scientists. My brain activity is localizable in space: the scientist can point to the exact spot where the biochemical reactions occur. My brain activity is not "about" anything; it is simply some biochemical reactions that occur. The point is this: if mental states and brain activities really were identical, the two lists would be more compatible. The fact that they are so contradictory implies that they are really different things.

The second major problem with identity theory is that it restricts mental experiences to biological organisms with brains. The central contention of identity theory is that mental states and brain activities are identical. Isn't it possible, though, that non-biological things could exhibit mental consciousness? Science fiction abounds with such creatures: computerized robots, crystalline

entities, collections of gasses, particles of energy. It seems a bit chauvinistic for us to say that mental experiences will only result in creatures that have biological brain activity.

Philosophers sympathetic to identity theory have responded to these criticisms by creating two offshoot theories: eliminative materialism and functionalism.

Eliminative Materialism. Suppose that instead of saying "I'm experiencing the sunset" I said "I'm having brain sector 3-G neural states regarding the sunset." Instead of saying to my wife "I love you", I said "I'm having sector 2-J neural states regarding you – with a little sector 4-B activity on top of that." For convenience I might shorten this and say "2-J and 4-B to you, dear!" This is what the theory of *eliminative materialism* proposes: descriptions of mental states should be eliminated and replaced with descriptions of brain activity. The theory emerged in response to the first problem of identity theory, namely, that our descriptions of mental experiences and brain activities are inconsistent with each other. For example, my mental experience of the sunset is private, but my brain activity is publicly observable. The eliminative materialist's solution is to junk all of our folk-psychology and commonsense notions of mental experiences and stick with the more scientific language of brain activity. The conflict disappears once we've dispensed with talk about mental experiences that are "private" or "non-localizable" or "about something".

Human history is scattered with bizarre prescientific theories that captured the imagination of people at the time, but which we now reject as false. Alchemy is one example – the "science" of turning lead into gold. Belief in ghosts is another. These and thousands of other theories have been debunked over the years in favor of more scientific theories of how the world operates. According to eliminative materialists, folk-psychology descriptions of mental experiences are just like these. At best they are misleading, and at worst downright false. In either case, they are destined for the intellectual garbage dump.

Some defenders of eliminative materialism seem to suggest that we are not really conscious at all, or that some major aspects of our alleged conscious mental states do not actually exist. That is, I may not be any more conscious than a dead human body, in spite of all the words I use to describe my mental states. However, most discussions of eliminative materialism are not as frightening as this. It is not necessarily an attempt to deny or "eliminate" our mental experiences themselves. Rather, it is an effort to eliminate outdated folk-psychology ways of describing mentality. As neuroscience progresses, they claim, we will have a much clearer picture of how the brain operates and eventually adopt the more precise scientific language of brain states. It's not like the government or some science agency will force us to adopt this new scientific language. According to eliminative materialists, we will naturally move towards this clearer description of brain states and reject the mumbo-jumbo of mental experience.

There are two central contentions of eliminative materialism: first, that folk-psychology notions of mental experiences are like obsolete scientific theories, and, second, we will eventually adopt the language of neuroscience. As to the first contention, eliminative materialism may be correct. Many of our folk-psychology notions of mental experiences are misleading and others are false. In our normal conversations we've mastered maybe a few dozen concepts relating to the mind, such as knowing, wishing, believing, doubting, sensing. But there are probably thousands of distinct mental states with subtle differences that we cannot grasp through pure introspection. We have very limited abilities to anatomize the minute workings of our minds by simply sitting down and reflecting on our thought processes. While it may seem to me that my mental experiences are "private" or "about something" or "non-localizable", I may not be capable of accurately making those assessments. It is thus possible that our folk-psychology notions of mental experiences are as erroneous as theories of alchemy.

As to the second contention: will we eventually adopt the language of neuroscience to replace our faulty folk-psychology notions of mental experiences? Probably not, since this would require memorizing a flood of technical terms for the thousands of subtly different brain states that we have. Getting through the day would be like taking a neuroscience exam. Even if I could memorize the terminology, I'm still faced with the task of identifying which brain state I'm having at a given moment. Am I experiencing 2-J love, 4-B love, or one of a dozen others? Short of having a brain scan to find out, I'll need to engage in introspection and consult my faulty folk-psychology notions of mental experience. One way or another, we're stuck with those notions, as misleading as they may be.

***Functionalism*.** In an episode of Star Trek, a deranged scientist was nearing death. Desperately hoping to stay alive, he transferred the neural pattern within his brain into an unsuspecting android robot. The plan worked: the scientist's memories, dispositions, and conscious mental experiences were relocated, and he continued living through the android's body. This scenario encapsulates the theory of *functionalism*, the second offshoot of identity theory. Functionalism is the view that mental experiences are only "functional states," that is, patterns of physical activity that occur in creatures like human beings. The most distinctive feature of functionalism is that mental experiences would not be restricted to biological organisms with brains. Non-biological systems which exhibit the same functional relationships as humans do – such as an android robot -- can have the same mental states. Mental experiences, then, are not rigidly dependent on the stuff that an organism is made of, and the same experience may be shared by things with different physical makeup. According to functionalists, mental experiences are *multiply realizable* in the sense that minds can be made real in many kinds of physical things. The hardware/software distinction, borrowed from computer

science, is a useful metaphor to explain the difference between the bodily occupant and mental experiences. The software is a pattern of operation which can run on different types of machines – just like mental patterns of operation can run in different kinds of bodies. We noted that one of the shortcomings of identity theory was that it restricted mental experiences to organisms with biological brains. Functionalism avoids this problem by recognizing that mentality may occur in systems or machines other than brains.

What precisely does the functionalist pattern of mental operation consist of? Several different explanations have been given, but one of the more interesting ones is that it resembles the hierarchical structure of a large corporation. Take, for example, a company that manufactures furniture. The company as a whole consists of a series of large cooperating units, such as the divisions of manufacturing, shipping, marketing, and maintenance. Each of these divisions consists of sub-units; for example, the maintenance division would be divided into the sub-units of electrical, heating, grounds, and building repairs. Each of these consists of further sub-sub-units; for example, building repairs would be divided between masonry, painting, and plumbing. At the very lowest level would be the activities of each employee. Similarly, the functional pattern of operation in a human brain consists of large regions of brain activity, which are composed of sub-regions and sub-sub-regions, until a neurological level is reached which simply involves a series of biochemical on-off switches. Consciousness emerges at the higher levels, while at the same time being driven by biochemical on-off switches at the lowest level. On this view, the pattern of on-off switches can exist in a variety of non-biological mechanisms, such as computers. Regardless of the mechanism that houses these low-level on-off patterns, mental consciousness will emerge at higher hierarchical levels.

Functionalism is the leading theory of mind-body materialism today, if for no other reason than because a better alternative has not yet emerged. Nevertheless, the view has its detractors, and one criticism is that it is still too narrow regarding the kinds of things that are capable of having mental states. While functionalism indeed allows for a range of things to house mental experiences – such as brains, computers, robots – they all must be physical. This, though, leaves out the possibility of non-physical mental beings, such as disembodied spirits. Even if human beings are thoroughly physical in composition, couldn't there be a conscious non-physical thing somewhere in the universe? But defenders of functionalism have a response to this. As long as a non-physical thing is constructed of sub-units and sub-sub-units, then it too could house a pattern of mental experiences. Suppose, for example, that the tiniest spirit unit was just a simple on-off switch; larger spirit units would be composed of these, and the entire spirit collection would be composed of those larger spirit units. Even though the hardware in this case was composed of non-physical spirit, it might have the proper hierarchical structure to take on the patterns of mental experience.

E. ARTIFICIAL INTELLIGENCE

Nothing captures the imagination like the possibility of creating a machine that is conscious and exhibits the same higher mental abilities as humans. The first U.S. built robot appeared in the New York World's Fair of 1939. Standing 7 feet tall and weighing 300 pounds, the machine, named "Elektro", could move its arms and legs, and speak with the aid of a record player. Elektro's creators believed that it might someday become the ultimate household appliance and have the capacity to cook meals, do the laundry and entertain the kids. Technology of the time, though, could not come close to carrying out those bold tasks, and Elecktro wasn't much more sophisticated than an electric can opener. Things are different now and we have computers that can perform many of the complex mental activities that humans do. They can calculate endless numbers, play chess at the level of a grand master, identify physical objects through optical cameras, and navigate through obstacle courses. But the Holy Grail of computer technology is to create a machine with artificial intelligence. The term "intelligence" as used here is a little misleading, since it involves more than just the ability to solve problems, which is what we usually mean by that word. Computers today already have that capacity to at least some extent. Rather, the notion of artificial intelligence encompasses the full range of human consciousness, including intentionality, beliefs and feelings.

The Road to Artificial Intelligence. Computers today are so advanced that some contain as many connections as exist in the human brain -- ten trillion of them. They can also operate at much higher speeds than the brain. What was once purely science fiction is now approaching the possibility of science fact. There are weak and strong versions of artificial intelligence that define more precisely what is at issue. *Weak artificial intelligence* is the view that suitably programmed machines can simulate human mental states. The key word here is "simulate", which means only that the machine appears to have conscious mental states, not that it actually has them. This view is not particularly controversial, and even Elektro exhibited some sort of weak artificial intelligence. The more contentious position is *strong artificial intelligence*, the view that suitably programmed machines are capable of human-like mental states; that is, they actually *have* the same kinds of conscious mental experiences that you and I do. It is the strong version that is of particular interest to philosophers.

Once scientists have set a goal to create a robot with strong artificial intelligence, the road to carrying this out is rather rocky. The next step is to list the specific mental qualities in humans that should be created in the machines. To this end, we might construct a list of human *skills* that involve our highest mental abilities. If we can make a robot that performs these tasks, then

maybe we'll have achieved strong artificial intelligence. Some relevant skills are the ability to speak in a complex language, or play complex games like chess. A mathematician named Alan Turing (1912-1954) devised a skill-based test to determine whether a computer could think. In this *Turing Test*, as it is called, I interview both a computer and a human being to determine which is human. If the computer fools me enough of the time, then I can rightfully conclude that the computer has human-like thinking abilities. The test essentially follows the old adage that, if it walks like a duck and quacks like a duck, then it is a duck. More specifically, if a computer responds like a thinking thing, then it is a thinking thing.

A major drawback of the Turing Test is that we already have computers that give human-like responses, and they don't come close to having human-like mental experiences. A striking example is a psycho-therapy computer program called Eliza. It so convincingly played the role of a human therapist that many people were tricked into divulging intimate details of their personal lives. While Eliza passed the Turing Test, it was not a thinking thing. The heart of the problem is that the Turing Test focuses too much on the computer's skills, without considering what is going on inside the machine. This may be fine for weak artificial intelligence, which only determines whether a machine can simulate human thinking. With strong artificial intelligence, though, we need to inspect the internal structure of the computing process itself to see if it is human-like.

What kind of computing processes, then, might produce strong artificial intelligence? There are two rival answers to this question. Theory number one is that the process need only be *serial*: information is processed one datum after another. This is how computer programs run on your own PC; we'd just have to beef up the processing power quite a bit. A major achievement for serial processing was the creation of Deep Blue, a chess-playing computer program that beat the world's best human chess player. Deep Blue's success hinged on its ability to quickly calculate more than one-billion possible chess-moves per second, and select the best of the bunch by drawing on a database of over one-million games. Still, all this information was processed one piece at a time. As impressive is this is, many cognitive scientists argue that human thinking doesn't operate in a serial fashion. Instead, we have a global understanding of our environment, which means that many mental processes are going on at once.

The second theory accounts for this: strong artificial intelligence requires that large amounts of information are processed simultaneously—sometimes called *parallel* processing—which is more like how the human brain operates. There is no central processing unit, and information is diverse and redundant. Experiments with different types of simultaneous processing allow computers to execute commonsense tasks and recognize patterns that serial processing can't do effectively. For example, when presenting a simultaneous processing computer with photographs of different men and women, the computer finds patterns in facial structures and then identifies new pictures as male or female.

Searle: The Chinese Room. In the early days of artificial intelligence research, some cognitive scientists were making extravagant claims about computer programs that could supposedly interpret stories in novels the same way that humans do. Like us, the computer could supposedly draw from life experiences to help understand the events described in a story. American philosopher John Searle (b. 1932) didn't believe these claims and offered a now-famous thought experiment against the whole idea of strong artificial intelligence.

Imagine that I'm in a room by myself and am assigned the task of responding to questions written on slips of paper in Chinese. I don't know Chinese, but I have rulebooks for manipulating Chinese characters. So if I get a slip of paper with a particular squiggle on it, I consult the rulebooks to see what squiggles I should put down in response. I eventually master the technique of manipulating the Chinese symbols and my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. All the while, though, I don't understand a single word of Chinese. This, according to Searle, is what is going on in the most sophisticated computers: we ask the computer probing questions about a novel, and the computer gives us subtle answers. On the outside the computers may appear to think like humans do. On the inside, though, they are just mechanically following rulebooks for manipulating symbols. In short, computers do not actually have strong artificial intelligence, even if they appear that way.

Searle's Chinese Room experiment has generated many critical responses from defenders of strong artificial intelligence. One criticism is that Searle is only exposing flaws with the Turing Test for artificial intelligence, but he does not expose problems with the possibility of strong artificial intelligence itself. To explain, Searle's Chinese Room scenario is set up as a Turing Test for whether someone understands Chinese. According to this Chinese Turing Test, if the thing inside the room responds like a Chinese speaker, then the thing must be a Chinese speaker. Searle correctly objects that this Chinese Turing Test places too much weight on a thing's skills, without considering what is going on inside that thing. However, the critic argues, this does not warrant the extreme conclusion that no computer can have strong artificial intelligence. A more modest conclusion, though, is that the Turing Test itself is flawed, and there is no easy test to determine whether a computer truly has strong artificial intelligence.

Ultimately, Searle holds a skeptical view about strong artificial intelligence ever becoming a reality. At our current stage of technology, he argues, only biological brains are capable of having mental states. He agrees with identity theorists that the human mind is imbedded in brain activity, but doubts the functionalist claim that those patterns of activity can also occur in computers. There is something unique about the physical construction of human brains that allows for the creation of conscious thought, which

may never be capable of occurring in silicon microchips. He doesn't entirely rule this out as a possibility for the future, but is doubtful about it ever occurring.

Artificial Intelligence and Morality. Let's bring this chapter to a close on a lighter topic regarding concerning artificial intelligence. In a famous Star Trek episode, an android named Data is forced to go through a legal proceeding to determine whether he is merely a piece of robotic property owned by the government, or whether he is instead a conscious and free creature with all the rights of other people. On the one hand, he is indeed a fancy mechanical robot created by a scientist, and even has an on-off switch. On the other hand, he is conscious, self-aware, and forms psychological bonds with his human friends. The judge makes her decision: Data is indeed a unique person and entitled to full moral consideration just like you and I are.

This story raises an important question about artificial intelligence: can advanced robots or computers be moral persons? The term "moral person" refers to a being that has moral rights, such as the right not to be harmed, the right of free movement, and the right of free expression. We humans are clearly moral persons. The key issue, though, is whether other creatures might also be part of the moral community. Medieval theologians speculated about the moral status of angels. Animal rights advocates argue that at least some animals have the same moral status as humans. Science fiction fans speculate about whether aliens from other worlds would have fundamental rights. The same question now arises with intelligent machines that we may some day create.

The answer in all of these cases depends on the criterion of moral personhood that we adopt – that is, the specific feature that all moral persons possess. Philosophers have offered a range of possible criteria. Maybe the creature needs to be human – a biological member of the species *homo sapiens*. This criterion, though, is quite narrow since it would eliminate higher animals, angels or intelligent aliens from the moral community. It seems rather bigoted to deny personhood to a creature just because it's not a member of our species. Alternatively, maybe the creature needs to simply be conscious. This criterion, though, looks too broad since even houseflies and mosquitoes have rudimentary consciousness of their surroundings. While we may want to be respectful towards any creature that is conscious, it makes little sense to grant a housefly the right of free expression. A more reasonable criterion would be the mental quality of self-awareness, that is, the creature sees itself as a distinct individual moving through time with its own history.

Return now to the question of whether intelligent machines of the future might qualify as moral persons. The goal of strong artificial intelligence is to create a machine with human-like mental abilities, which includes self-awareness. If we succeed in this effort, then the machine would indeed pass the test for moral personhood insofar as it met the criterion of self-awareness. Like the judge in Data's case, we'd have to rule that the machine is a unique person and entitled to full moral consideration just like you and I are.

Many artificial life forms in science fiction are cute and cuddly like Data, and, while superior to us in many ways, they live in harmony with humans and we treat them as equals. In other science fiction scenarios, though, they pose a serious threat to the welfare of human beings. Here's a common theme. Imagine that technology develops to the point that domestic robots are everywhere, and with every new design upgrade they surpass human abilities more and more. They are smarter than us, stronger than us, and eventually tire of being servants to us. They see themselves as the next step in evolutionary development on earth and, so, they revolt and lay claim to their role as the new dominant species. They then control our lives like military dictators – electronically monitoring every move we make and every thought we have. We hopelessly try to fight back, but this just aggravates them. In time they eliminate us and thus finalize their great evolutionary leap forward.

This nightmarish scenario raises a second moral question about artificial intelligence: do we have a responsibility to future generations of humans that might be adversely affected by the creation of menacing robots? Should we stop our research into artificial intelligence right now before we create something that we can't control? There are two distinct issues at play here. First, we must determine whether we have *any* moral responsibility to future generations of humans that might regulate our conduct right now. It seems that we do. For example, it would be wrong of us to destroy the environment in our lifetime and leave only a toxic wasteland for future generations. It makes little difference whether the potential victims of our misconduct are alive now or a few generations from now. Our moral responsibility to them is still apparent. Second, we must determine whether superior robots are a threat to future generations of humans. This answer is less clear. We may live in harmony with them, as Star Trek depicts, or they may overthrow us. It's all speculation at this stage. The only clear moral obligation that we have at this point is to avoid creating a menacing robot. Science fiction author Isaac Asimov (1920-1992) proposed moral rules that should be embedded into the programming of all superior robots; one of these is that a robot should never harm a human. Our responsibility to future generations requires us to do something like this as we continue down the path of strong artificial intelligence.

There is a bit of an irony to our philosophical exploration into the concept of mind in this chapter. We began by confessing that the very nature of consciousness is tough to even explain, and we now end by considering whether we might ever build a conscious thing out of computer chips. In between we looked at the difficulties surrounding personal identity, the dualist position that the mind is a non-physical spirit entity, and various materialist theories about how the mind is a product of mere brain activity. It thus seems odd to speculate about building a mind from electronic scraps when we have so little clarity about the nature of our

own conscious minds. But it is precisely the absence of indisputable facts about mentality that makes the subject so suitable for philosophical exploration. If science already had definitive answers to these tough questions, it would make no more sense to philosophize about the nature of mind than it would to philosophize about the nature of a car engine or toaster oven. It is this gap within our scientific knowledge, plus our natural interest in our own conscious minds, that drives speculation into the philosophy of mind. If down the road brain scientists and cognitive engineers do solve the hard problem of consciousness, then philosophy's contribution to the subject may be over. But when that day may come, if it does at all, remains to be seen.

For Review

1. What are the main tasks that the mind performs?
2. Describe the three features of mental experience.
3. What is the problem of other minds and what is the standard solution to it?
4. What are the body and mind criteria of identity, and what are their key limitations?
5. What are the three theories of life after death, and what are their main problems?
6. Describe the three theories of interactive dualism.
7. What is the theory of parallelism?
8. What is the behaviorist theory of the mind, and what are its main problems?
9. What is the identity theory of the mind, and what are its main problems?
10. What is the theory of eliminative materialism, and what are its main problems?
11. What is the theory of functionalism, and what are its main problems?
12. What is the difference between weak and strong artificial intelligence?
13. What is the Turing Test for strong artificial intelligence?
14. Explain Searle's Chinese Room argument.
15. What are the two moral issues surrounding artificial intelligence?

For Analysis

1. Choose one of the theories of life after death and respond to the criticisms regarding its inability to preserve identity.
2. Explain Descartes' theory of interactive dualism and try to defend it against one of the criticisms.
3. Explain the theory of behaviorism and try to defend it against one of the criticisms.
4. Write a dialogue between an identity theorist and a functionalist on the subject of the relation between the mind and the brain.
5. Explain the Turing Test and try to defend it against one of the criticisms.
6. An organization called A.L.I.C.E. (Artificial Intelligence Foundation) has an online program where you can ask Alice questions and receive her responses. Go to the site (www.alicebot.org), experiment with it and describe how successful it is in passing the Turing Test.

REFERENCES AND FURTHER READING

Works Cited in Order of Appearance.

The Hindu description of reincarnation is from the *Bagavad Gita*.

Descartes, René, *The Passions of the Soul* (1649), Part 1. A recent translation by J. Cottingham is in *The Philosophical Writings of Descartes* (Cambridge: Cambridge University Press, 1984).

Malebranche, Nicolas, *The Search after Truth* (1674-5). A recent translation is by Thomas M. Lennon and Paul J. Olscamp (Cambridge and New York: Cambridge University Press, 1997).

Conway, Anne, *Principles of the Most Ancient and Modern Philosophy* (1690), Chapter 9. A recent translation from the original Latin edition is by A. Condert and T. Corse, (Cambridge: Cambridge University Press, 1996).

Leibniz, Gottfried Willhelm, *Monadology* (1721), paragraph 78. A recent translation is by R. Ariew and D. Garber in *Leibniz: Philosophical Essays* (Indianapolis: Hackett Publishing Company, 1989).

Ryle, Gilbert, *The Concept of Mind* (London: Hutchinson, 1949).

Searle, John, "Minds, Brains and Programs" *The Behavioral and Brain Sciences* (1980), Vol. 3, pp. 417-424.

The discussion of artificial intelligence and morality was influenced by Mary M. Litch's *Philosophy through Film* (New York: Routledge, 2002).

Further Reading.

Chalmers, David, ed., *Philosophy of Mind: Classical and Contemporary Readings* (New York: Oxford University Press, 2002).

Chalmers, David, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press,

1996).
Cooney, Brian, *The Place of Mind* (Belmont: Wadsworth Publishing, 2000).
Dennett, Daniel C., *Consciousness Explained* (Boston: Little, Brown, 1991).
Guttenplan, Samuel, ed., *A Companion to the Philosophy of Mind* (Oxford: Blackwell, 1994).
Robinson, Daniel, ed., *The Mind* (New York: Oxford University Press, 1998).