The Behavior Analyst 2012, 35, 1–16 No. 1 (Spring)

Making IBM's Computer, Watson, Human

Howard Rachlin
Stony Brook University

This essay uses the recent victory of an IBM computer (Watson) in the TV game, *Jeopardy,* to speculate on the abilities Watson would need, in addition to those it has, to be human. The essay's basic premise is that to be human is to behave as humans behave and to function in society as humans function. Alternatives to this premise are considered and rejected. The viewpoint of the essay is that of teleological behaviorism. Mental states are defined as temporally extended patterns of overt behavior. From this viewpoint (although Watson does not currently have them), essential human attributes such as consciousness, the ability to love, to feel pain, to sense, to perceive, and to imagine may all be possessed by a computer. Most crucially, a computer may possess self-control and may act altruistically. However, the computer's appearance, its ability to make specific movements, its possession of particular internal structures (e.g., whether those structures are organic or inorganic), and the presence of any nonmaterial ''self,'' are all incidental to its humanity.

Key words: human nature, IBM, Watson, pain, rationality, self-control, social cooperation, Turing test, vitalism

Recently an IBM computer named Watson defeated two champions of the TV quiz show, *Jeopardy*. Excelling at *Jeopardy* requires understanding natural language questions posed in the form of ''answers,'' for example, ''He was the last president of the Soviet Union.'' The correct ''question'' is, ''Who is Mikhail Gorbachev?'' Watson consists of numerous high-power computers that operate in parallel to search a vast self-contained database (no Web connection). The computers fill a room. The only stimuli that affect Watson are the words spoken into its microphone, typed into its keyboard, or otherwise fed into it; the only action Watson is capable of is to speak or print out its verbal response. Information in and information out.

According to the IBM website, Watson's amazing performance in the quiz game requires ''natural language processing, machine learning, knowledge representation and reasoning, and deep analytics.'' Still, it is clear, Watson is not human. In considering what might make Watson human I hope to throw some light on the question, What makes us human? What are the minimal requirements of humanity?

I will call Watson, so modified, Watson II. Many people believe that nothing whatsoever could make a machine such as Watson human. Some feel that it is offensive to humanity to even imagine such a thing. But I believe it is not only imaginable but possible with the technology we have now. Because in English we recognize the humanity of a person by referring to that person as ''he'' or ''she'' rather than ''it'' (and because the real Watson, IBM's founder, was a man), I will refer to Watson as ''it'' and Watson II as ''he.'' But this is not to imply that making Watson human would be a good thing or a profitable thing for IBM to do. As things stand, Watson has no interests that are different from IBM's interests. Because its nonhumanity allows IBM to own it as a piece of property, Watson may be exploited with a clear conscience. But that does not mean that it is

Preparation of this article was supported by Grant DA02652021 from the National Institute of Drug Abuse. The content is solely the responsibility of the author and does not necessarily represent the views of the National Institutes of Health.

Address correspondence to Howard Rachlin, Psychology Department, Stony Brook University, Stony Brook, New York 11794 (e-mail: howard.rachlin@sunysb.edu).

pointless to speculate about what would make Watson human. Doing so gives us a place to stand as we contemplate what exactly makes *us* human, a prolegomenon for any discussion of moral questions.

CAN A COMPUTER EVER BE CONSCIOUS?

To put this question into perspective, let me repeat an anecdote from an earlier work (Rachlin, 1994, pp. 16–17). Once, after a talk I gave, a prominent philosopher in the audience asked me to suppose I were single and one day met the woman of my dreams (Dolly), who was beautiful, brilliant, witty, and totally infatuated with me. We go on a date and I have the greatest time of my life. But then, the next morning, she reveals to me that she is not a human but a robot, silicon rather than flesh and blood. Would I be disappointed? the philosopher wanted to know. I admitted that I would be disappointed. She was just going through the motions, her confession would have implied. She did not really have any feelings. Had I been quicker on my feet, however, I would have answered him something like this: Imagine another robot, Dolly II, an improved model. This robot, as beautiful, as witty, as sexually satisfying as Dolly I, doesn't reveal to me, the next day, that she's a machine; she keeps it a secret. We go out together for a month and then we get married. We have two lovely children (half-doll, half-human) and live a perfectly happy life together. Dolly II never slips up. She never acts in any way but as a loving human being, aging to all external appearances as real human beings do (but gracefully), retaining her beauty, her wit, her charm. At this point she reveals to me that she is a robot. Would the knowledge that the chemistry of her insides was inorganic rather than organic make any difference to me or her loving children or her friends? I don't think so. The story of Dolly II reveals that the thing that wasn't there in Dolly I (the soul, the true love) consists of *more* behavior.

Some philosophers would claim that Dolly II cannot be a real human being. According to them, she would be essentially a zombie. Lacking human essence, she would have "the intelligence of a toaster" (Block, 1981, p. 21); presumably we could treat her with no more consideration for her feelings than in dealing with a toaster. Because we cannot perceive the inner states of other people, such an attitude poses a clear moral danger. If the humanity of others were an essentially nonphysical property within them, there would be no way of knowing for sure that others possess such a property; it may then be convenient for us to suppose that, despite their "mere" behavior, they do not. But, although the teleological behaviorism I espouse can be defended on moral grounds (Baum, 2005), the present essay relies on functional arguments. A behavioral conception of humanity is better than a spiritual or neurocognitive conception, not because it is more moral but, as I shall try to show, because it is potentially more useful.[1]

For a behaviorist, consciousness, like perception, attention, memory, and other mental activities, is itself *not* an internal event at all. It is a word we use to refer to the organization of long-term behavioral patterns as they are going on. Consider an orchestra playing Beethoven's Fifth symphony. At any given moment a violinist and an oboist may both be sitting quite still with their instruments at their

[1] I do not claim that behaviorism is free of moral issues. Many people, particularly infants and the mentally disabled, lack essential human behavioral characteristics. A behavioral morality, when it is developed, needs to accommodate the humanity of these people, perhaps in terms of their former or expected behavior or their role in human social networks but never in terms of internal, essentially unobservable actions or states.

sides. Yet they are both in the midst of playing the symphony. Because the violinist and oboist are playing different parts, we can say that, despite their identical current actions (or nonactions), their mental states at that moment are different. The teleological behaviorist does not deny that the two musicians have mental states or that these states differ between them. Moreover, there must be differing internal mechanisms that underlie these states. But the mental states themselves *are* the musicians' current (identical) actions in their (differing) patterns. A behaviorist therefore need not deny the existence of mental states. For a teleological behaviorist such as myself, they are the main object of study (Rachlin, 1994). Rather, for a teleological behaviorist, mental states such as perception, memory, attention, and the conscious versions of these states are themselves temporally extended patterns of behavior. Thus, for a teleological behaviorist, a computer, if it behaves like a conscious person, would be conscious. If Watson could be redesigned to interact with the world in all essential respects as humans interact with the world, then Watson would be human (he would be Watson II) and would be capable of behaving consciously.

Are Our Minds Inside of Us?

If, like most modern scientists and philosophers, you believe that no other forces than common physical and chemical ones are active within the organism, and you also believe that our minds must be inside of us, you will be led to identify consciousness with activity in the brain. Thus, current materialist studies of consciousness are studies of the operation of the brain (e.g., Ramachandran, 2011). Neurocognitive theories may focus on the contribution to consciousness of specific brain areas or groups of neurons. Or they may be more broadly based, attributing conscious thought to the integration of stimulation over large areas of the brain. An example of the latter is the work of Gerald Edelman and colleagues (Tononi & Edelman, 1998) in which consciousness, defined in behavioral terms, is found to correlate with the occurrence of reciprocal action between distributed activity in the cortex and thalamus. Edelman's conception of consciousness, like that of teleological behaviorism, is both behavioral and molar, but it is molar *within* the nervous system. This research is interesting and valuable. As it progresses, we will come closer and closer to identifying the internal mechanism that underlies human consciousness. Some day it may be shown that Edelman's mechanism is sufficient to generate conscious behavior. But it seems highly unlikely to me that such a mechanism, or any particular mechanism, will ever be shown to be *necessary* for consciousness. If it were possible to generate the same behavior with a different mechanism, that behavior would be no less conscious than is our behavior now. Why? Because consciousness is in the behavior, not the mechanism.

There is currently a movement in the philosophy of mind called *enacted mind* or *extended cognition*. This movement bears some resemblances to behaviorism. According to the philosopher Alva Noë (2009), for example, the mind is not the brain or part of the brain and cannot be understood except in terms of the interaction of a whole organism with the external environment. Nevertheless, for these philosophers, the brain remains an important component of consciousness. They retain an essentially neurocognitive view of the mind while expanding its reach spatially, beyond the brain, into the peripheral nervous system and the external environment.

For a behaviorist, it is not self-evident that our minds are inside of us at all. For a teleological behaviorist, all mental states (including sensations,

perceptions, beliefs, knowledge, even pain) are rather patterns of overt behavior (Rachlin, 1985, 2000). From a teleological behavioral viewpoint, consciousness is not the organization of neural complexity, in which neural activity is distributed widely over space in the brain, but the organization of behavioral complexity in which overt behavior is distributed widely over time. To study the former is to study the mechanism that underlies consciousness, as Edelman and his colleagues are doing, but to study the latter is to study consciousness itself.

Widespread organization is characteristic of much human behavior. As such, it must have evolved either by biological evolution over generations or by behavioral evolution within the person's lifetime. That is, it must be beneficial for us in some way. The first question the behaviorist asks is therefore, "W*hy* are we conscious?" Nonhumans, like humans, may increase reward by patterning their behavior over wider temporal extents (Rachlin, 1995) and may learn to favor one pattern over another when it leads to better consequences (Grunow & Neuringer, 2002). The difference between humans and nonhumans is in the temporal extent of the patterns.[2]

The place to start in making Watson human is not at appearance or movement but at human *function* in a human environment. Let us therefore concede to Watson only the degree of movement necessary to speak and to understand speech. Like Stephen Hawking, Watson II will have a grave disability but be no less human for that. We ask, What might Watson II want and how might he manage to get it?

WATSON'S FUNCTION IN HUMAN SOCIETY

Watson already has a function in human society; it has provided entertainment for hundreds of thousands of people. But Watson would quickly lose entertainment value if it just continued to play *Jeopardy* and kept winning. There is talk of adapting Watson to remember the latest medical advances, to process information on the health of specific individuals, and to answer medical questions. Other functions in business, law, and engineering are conceivable. In return, IBM, Watson's creator, provides it with electrical power, repairs, maintenance, continuous attention, and modification so as to better serve these functions. Moreover, there is a positive relation between Watson's work and Watson's payoff. The more useful Watson is to IBM, the more IBM will invest in Watson's future and in future Watsons. I do not know what the arrangement was between *Jeopardy*'s producers and IBM, but if some portion of the winnings went to Watson's own maintenance, this would be a step in the right direction. It is important to note that this is the proper direction. To make Watson human, we need to work on Watson's function in society. Only after we have determined Watson II's minimal functional requirements should we ask how those requirements will be satisfied.

Functions in society, over and above entertainment, are necessary if we are to treat Watson II as human. Let us assume that Watson is given such functions.[3] Still they are far from sufficient to convince us to treat Watson as human. A further requirement is that Watson's reasoning be modified to better approximate human reasoning.

[2] The CEO of a company is rewarded not for anything he or she does over an hour or a day or a week but for patterns in his or her behavior extended over months and years. Despite this, it has been claimed that corporate executives' rewards are still too narrowly based. Family-owned businesses measure success over generations and thus may be less likely than corporations to adopt policies that sacrifice long-term for short-term gains.

[3] IBM is doing just that. It is currently developing and marketing versions of Watson for law, medicine, and industry (Moyer, 2011).

Watson's Memory And Logic

A human quality currently lacking in Watson's logic is the ability to take its own weaknesses into account in making decisions. Here is an example. You are driving and come to a crossroads with a traffic light that has just turned red. You are in a moderate hurry but this is no emergency. You have a clear view in all four directions. There is no other car in sight and no policeman in sight. The odds of having an accident or getting a ticket are virtually zero. Should you drive through the red light? Some of us would drive through the light, but many of us would stop anyway and wait until it turned green. Why? Let us eliminate some obvious answers. Assume that the costs of looking for policemen and other cars are minimal. Assume that you are not religious so that you do not believe that God is watching you. Assume that you are not a rigid adherent to all laws regardless of their justifiability. Then why stop? One reason to stop is that you have learned over the years that your perception and judgment in these sorts of situations are faulty. You realize that, especially when you are in a hurry, both your perception and reasoning tend to be biased in favor of quickly getting where you're going. To combat this tendency you develop the personal rule: Stop at all red lights (unless the emergency is dire or unless the light stays red so long that it is clearly broken). This rule, as you have also learned, serves you well over the long run.

Here is another case. You are a recovering alcoholic. You have not taken a drink for a full year. You are at a party and trying to impress someone. You know that having one or two drinks will cause you no harm and significantly improve your chances of impressing that person. You know also that you are capable of stopping after two drinks; after all, you haven't taken a drink in a year. Why not have the drink? No reason, Watson would say, unless the programmers had arbitrarily inserted the rule, "Never drink at parties." (But that would make Watson still more machine-like.) To be human, Watson would have to itself establish the rule and override its own logical mechanism, because it knows its own calculations are faulty in certain situations. Watson does not have this kind of logic. It does not need it currently. But it would need such rules if it had to balance immediate needs with longer term needs, and those with still longer term needs, and so forth. As it stands, Watson can learn from experience, but its learning is time independent.[4]

Watson, I assume, obeys the economic maxim: Ignore sunk costs. It can calculate the benefits of a course of action based on estimated returns from that action from now to infinity. But it will not do something just because that is what it has done before. As long as its perception and logic capacity remain constant, Watson will not now simply follow some preconceived course of action. But, if its perception and logic mechanism will be predictably weaker at some future time, Watson will be better off deciding on a plan and just sticking to it than evaluating every alternative strictly on a best estimate of that alternative's own merits. Hal, the computer of *2001: A Space Odyssey,* might have known that its judgment would be impaired if it had to admit that it made a wrong prediction; Hal should have trusted its human operators to know better than it did at such a time. But Hal was machine-like in its reliance on its own logic.

Paying attention to sunk costs often gets people into trouble. It is, after all, called a fallacy by economists. Because

[4] Here is an example of the usefulness of the behavioristic approach. Addiction may be seen as a breakdown of temporally extended behavioral patterns into temporally narrower patterns. It would then be directly accessible to observation and control (Rachlin, 2000).

they had invested so much money in its development, the British and French governments stuck with the Concorde supersonic airplane long after it had become unprofitable. People often hold stocks long after they should have sold them or continue to invest in personal relationships long after they have become painful. Their reasons for doing so may be labeled "sentimental." But we would not have such sentimental tendencies if they were not generally useful, however disastrous they may turn out in any particular case. Our tendency to stick to a particular behavioral pattern, no matter what, can get out of hand, as when we develop compulsions. But, like many other psychological malfunctions, compulsiveness is based on generally useful behavioral tendencies.

A satisfactory answer to a *why* question about a given act may be phrased in terms of the larger pattern into which the act fits. *Why* is he building a floor? Because he is building a house. For Watson to be human, we must be able to assign reasons (i.e., functional explanations) for what he does: Q. Why did Watson bet $1,000 on a *Jeopardy* daily double? A. Because that maximizes the expected value of the outcome on that question. Q. Why maximize the expected value of the outcome? A. Because that improves the chance of winning the game.[5] Q. Why improve the chance of winning the game? A. Because that will please IBM. Q. Why please IBM? A. Because then IBM will maintain and develop Watson and supply it with power. Thus, Watson may have longer term goals and shorter term subgoals. This is all it needs to have what philosopher's call *intentionality*. Philosophers might say, but Watson does not know "what it is like" to have these goals, whereas humans do know "what it is like."

Do we know what it is like to be our brothers, sisters, mothers, fathers, any better than we know what it is like to be a bat (Nagel, 1974)? Not if "what it is like" is thought to be some ineffable physical or nonphysical state of our nervous systems hidden forever from the observations of others. The correct answer to "What is it like to be a bat?" is "to behave, over an extended time period, and in a social context, as a bat behaves." The correct answer to "What is it like to be a human being?" is "to behave, over an extended time period, and in a social context, as a human being behaves."

Will Watson II Perceive?

Watson can detect minuscule variations in its input and, up to a point, can understand their meaning in terms of English sentences and their relation to other English sentences. Moreover, the sentences it detects are directly related to what is currently its primary need (its reason for being), which is coming up quickly with the right answer ("question") to the *Jeopardy* question ("answer") and learning by its mistakes to further refine its discrimination. Watson's perception of its electronic input, however efficient and refined, is also very constrained. That is because of the narrowness of its primary need (to answer questions).

But Watson currently has other needs: a steady supply of electric power with elaborate surge protection, periodic maintenance, a specific temperature range, protection from the elements, protection from damage or theft of its hardware and software. I assume that currently there exist sensors, internal and external, that monitor the state of the systems that supply these needs. Some of the sensors are probably external to the machine itself, but let us imagine all are located on Watson II's surface.

[5] This is as far as Watson can go itself. The subsequent reasons may be assigned solely to its human handlers. But Watson II, the human version of Watson, would have to have the subsequent reasons in himself.

Watson currently has no way to satisfy its needs by its own behavior, but it can be given such powers. Watson II will be able to monitor and analyze his own power supply. He will have distance sensors to monitor the condition of his surroundings (his external environment) and detect and analyze movement (his speech) in the environment in terms of benefits and threats to his primary and secondary functions. He will be able to organize his speech into patterns, those that lead to better functioning in the future and those that lead to worse. He will be able to act in such a way that benefits are maximized and threats are minimized by his actions. That is, he will be able to discriminate among (i.e., behave differently in the presence of) different complex situations that may be extended in time. He will be able to discriminate one from another of his handlers. He will be able to discriminate between a person who is happy and a person who is sad. He will be able to discriminate between a person who is just acting happy and one who truly is happy, between a person with hostile intentions towards him and a person with good intentions. I do not believe that a system that can learn to make such discriminations is beyond current technology. The most sophisticated, modern poker-playing computers currently bluff and guess at bluffs depending on the characteristics of specific opponents. These very subtle discriminations would extend to Watson II's social environment. Watson II will have the power to lie and to deceive people. Like other humans, he will have to balance the immediate advantages of a lie with the long-term advantage of having a reputation for truth telling and the danger of damaging a person's interests that might overlap with his own interests. These advantages may be so ineffable that he may develop the rule: Don't lie except in obvious emergencies. Like the rule discussed above (stop at red lights except in obvious emergencies), this might serve well to avoid difficult and complex calculations and free up resources for other purposes.

With such powers, Watson II's perception will function for him as ours does for us. It will help him in his social interactions as well as his interactions with physical objects that make a difference in satisfying his needs. What counts for Watson II's humanity is not *what* he perceives, not even *how* he perceives, but *why* he perceives; his perception must function in the same way as ours does.

Will Watson II Be Able to Imagine?

Watson seems to have a primitive sort of imagination. Like any computer, it has internal representations of its input. But a picture in your head, or a coded representation of a picture in your head, although it may be part of an imagination mechanism, is not imagination itself. Imagination itself is behavior; that is, acting in the absence of some state of affairs as you would in its presence. Such behavior has an important function in human life; that is, to make perception possible. Pictures in our heads do not themselves have this function.

To illustrate the distinction, suppose two people, a woman and a man, are asked to imagine a lion in the room. The woman closes her eyes, nods, says, "Yes, I see it. It has a tail and a mane. It is walking through the jungle." The man runs screaming from the room. The man is imagining an actual lion. The woman would be imagining not a lion but a picture of a lion; in the absence of the picture, she is doing what she would do in its presence. An actor on a stage is thus a true imaginer, and good acting is good imagination, not because of any picture in the actor's head but because he or she is behaving in the absence of a set of conditions as he or she would if they were actually present.

What would Watson need to imagine in this sense? Watson already has

this ability to a degree. As a question is fed in, Watson does not wait for it to be completed. Watson is already guessing at possible completions and looking up answers. Similarly, we step confidently onto an unseen but imagined floor when we walk into a room. The outfielder runs to where the ball will be on the basis of the sound of the bat and a fraction of a second of its initial flight. We confide in a friend and do not confide in strangers on the basis of their voices on the telephone. On the basis of small cues, we assume that perfect strangers will either cooperate with us in mutual tasks or behave strictly in their own interests. Of course we are often wrong. But we learn by experience to refine these perceptions. This tendency, so necessary for everyday human life, to discriminate on the basis of partial information and past experience, and to refine such discriminations based on their outcomes vis-à-vis our needs, will be possessed by Watson II.

Will Watson II Feel Pain?

In a classic article, Dennett (1978) took up the question of whether a computer could ever feel pain. Dennett designed a pain program that duplicated in all relevant respects what was then known about the human neural pain processing system and imagined these located inside a robot capable of primitive verbal and nonverbal behavior. But at the end, he admitted that most people would not regard the robot as actually in pain. I agree. To see what the problem is, let us consider the Turing test, invented by the mathematician Alan Turing (1912–1954), as a means for determining the degree to which a machine can duplicate human behavior.

The problem with the Turing test. Imagine the machine in question and a real human side by side behind a screen. For each there is an input device (say, a computer keyboard) and an output device (say, a computer screen) by which observers may ask questions and receive answers, make comments, receive comments, and generally communicate. The observer does not know which inputs and outputs are going to and coming from the human and which are going to and coming from the machine. If, after varying the questions over a wide range so that, in the opinion of the observer, only a real human can meaningfully answer them, the observer still cannot reliably tell which is the computer and which is the machine, then, within the constraints imposed by the range and variety of the questions, the machine is human, regardless of the mechanism by which the computer does its job.

What counts is the machine's behavior, not the mechanism that produced the behavior. Nevertheless there is a serious problem with the Turing test; it ignores the function of the supposedly human behavior in human society. Let us agree that Dennett's (1978) computer would pass the Turing test for a person in pain. Whatever questions or comments typed on the computer's keyboard, the computer's answers would be no less human than those of the real person in real pain at the time. Yet the machine is clearly not really in pain, but the person is.[6]

The Turing test is a behavioral test. But as it is typically presented, it is much too limited. If we expand the

[6] For Dennett, the lesson of the Turing test for pain is that certain mental states, such as pain, pleasure, and sensation, that philosophers call "raw feels," are truly private and available only to introspection. That, Dennett believes, is why the Turing test fails with them, not, as I believe, because it cannot capture their social function. Dennett believes that other mental states, called "propositional attitudes," such as knowledge, thought, reasoning, and memory, unlike raw feels, may indeed be detectable in a machine by means of the Turing test. But, like raw feels, propositional attitudes have social functions. Thus, the Turing test is not adequate to detect either kind of mental state in a machine. Watson may pass the Turing test for logical thinking with flying colors, but unless the actual function of its logic is expanded, in ways discussed here, Watson's thought will differ essentially from human thought.

test by removing the screen and allowing the observer to interact with the mechanisms in meaningful ways over long periods of time (say, in games that involve trust and cooperation), and the computer passed the test, we would be approaching the example of Dolly II. Such a Turing test would indeed be valid. Let us call it the *tough* Turing test.

Passing the tough Turing test for pain. In an article on pain (Rachlin, 1985), I claimed that a wagon with a squeaky wheel is more like a machine in pain than Dennett's computer would be (although of course the wagon is not really in pain either). What makes the wagon's squeak analogous to pain? The answer lies in how we interact with wagons as opposed to how we interact with computers. The wagon clearly needs something (oil) to continue to perform its function for our benefit. It currently lacks that something and, if it does not get it soon, may suffer permanent damage, may eventually, irreparably, lose its function altogether. The wagon expresses that need in terms of a loud and annoying sound that will stop when the need is satisfied. "You help me and I'll help you," it seems to be saying. "I'll help you in two ways," the wagon says. "First, I'll stop this annoying sound; second, I'll be better able to perform my function."

To genuinely feel pain, Watson must interact with humans in a way similar to a person in pain. For this, Watson would need a system of lights and sounds (speech-like if not speech itself), the intensity of which varied with the degree of damage and the quality of which indicated the nature of the damage. To interact with humans in a human way, Watson would need the ability to recognize individual people. Currently Watson has many systems operating in parallel. If it became a general purpose adviser for medical or legal or engineering problems, it might eventually need a system for allocating resources, for acknowledging a debt to individual handlers who helped it, a way of paying that debt (say, by devoting more of its resources to those individuals, or to someone they delegate) and, correspondingly, to punish someone who harmed it. In addition, Watson II would be proactive, helping individuals on speculation, so to speak, in hope of future payback. Furthermore, Watson II would be programmed to respond to the pain of others with advice from his store of diagnostic and treatment information and with the ability to summon further help (e.g., to dial 911 and call an ambulance). In other words, to really feel pain, Watson would need to interact in an interpersonal economy, giving and receiving help, learning whom to trust and whom not to trust, responding to overall rates of events as well as to individual events. In behaving probabilistically, Watson II will often be "wrong," too trusting or too suspicious. But his learning capacity would bring such incidents down to a minimal level.

Watson II will perceive his environment in the sense discussed previously; learning to identify threats to his well-being and refining that identification over time. He will respond to such threats by signaling his trusted handlers to help remove them, even when the actual damage is far in the future or only vaguely anticipated. Again, the signals could be particular arrangements of lights and sounds; in terms of human language, they would be the vocabulary of fear and anxiety. In situations of great danger, the lights and sounds would be bright, loud, continuous, annoying, and directed at those most able to help. The first reinforcement for help delivered would be the ceasing of these annoyances (technically, negative reinforcement). But, just as the wagon with the squeaky wheel functions better after oiling, the primary way that Watson II will reinforce the responses of his social circle will be his better functioning.

It is important to note that "Is Watson really in pain?" cannot be separated from the question, "Is Watson human?" A machine that had the human capacity to feel pain, and only that capacity, that was not human (or animal) in any other way, could not really be in pain. A Watson that was not a human being (or animal) could not really be in pain. And this is the case for any other individual human trait. Watson could not remember, perceive, see, think, know, or believe in isolation. These are human qualities by definition. Therefore, although we consider them one by one, humanity is a matter of all (or most) or none. Watson needs to feel pain to be human but also needs to be human before it can feel pain. But, a Watson that is human (i.e., Watson II) in other respects *and* exhibits pain behavior, as specified above, would really be in pain.

What do we talk about when we talk about pain? Once we agree that the word *pain* is not a label for a nonphysical entity within us, we can focus on its semantics. Is it efficient to just transfer our conception of pain (its meaning) from a private undefined spiritual experience that no one else can observe to a private and equally undefined stimulus or response that no one else can observe? Such a shift simply assumes the privacy of pain (we all know it; it is intuitively obvious) and diverts attention from an important question: What is gained by insisting on the privacy of pain? Let us consider that question.

In its primitive state, pain behavior is the unconditional response to injury. Sometimes an injury is as clear to an observer as is the behavior itself; but sometimes the injury may be internal, such as a tooth cavity, or may be abstract in nature, such as being rejected by a loved one. In these cases the help or comfort we offer to the victim cannot be contingent on a detailed checking out of the facts; there is no time; there may be no way to check. Our help, to be effective, needs to be immediate. So, instead of checking out the facts, we give the person in pain, and so should we give Watson II, the benefit of the doubt. Like the fireman responding to an alarm, we just assume that our help is needed. This policy, like the fireman's policy, will result in some false alarms. Watson II's false alarms might teach others to ignore his signals. But, like any of us, Watson II would have to learn to ration his pain behavior, to reserve it for real emergencies. Watson II's balance point, like our balance points, will depend on the mores of his society. That society might (like the fire department) find it to be efficient to respond to any pain signals regardless of the frequency of false alarms; or that society, like a platoon leader in the midst of battle, might generally ignore less than extreme pain signals, ones without obvious causative damage.

Our internalization of pain thus creates the risk that other people's responses to our pain will, absent any injury, reinforce that pain. This is a risk that most societies are willing to take for the benefit of quickness of response. So, instead of laying out a complex set of conditions for responding to pain behavior, we imagine that pain is essentially private, that only the person with the injury can observe the pain. We take pain on faith. This is a useful shortcut for everyday-life use of the language of pain (as it is for much of our mentalistic language), but it is harmful for the psychological study of pain and irrelevant to the question of whether Watson II can possibly be in pain. Once Watson II exhibits pain behavior and other human behavior in all of its essential respects, we must acknowledge that he really is in pain. It may become convenient to suppose that his pain is a private internal event. But this would be a useful convenience of everyday linguistic communication, like the convenience of supposing that the sun revolves around the earth, not a statement of fact.

Because Watson II's environment would not be absolutely constant, he would have to learn to vary his behavior to adjust to environmental change. *Biological evolution* is the process by which we organisms adjust over generations, in structure and innate behavioral patterns, as the environment changes. *Behavioral evolution* is the process by which we learn new patterns and adjust them to environmental change within our lifetimes. In other words, our behavior evolves over our lifetimes just as our physical structure evolves over generations.[7] We hunt through our environment for signals that this adaptation is working. Similarly, Watson II will adjust its output in accordance with complex and temporally extended environmental demands. Watson II will signal that things are going well in this respect with lights and sounds that are pleasurable to us. With these signals he will not be asking his handlers for immediate attention, as he would with pain signals. But such pleasing signals will give Watson II's handlers an immediate as well as a long-term incentive to reciprocate, to give Watson II pleasure and avoid giving him pain.

[7] I say *behavioral evolution,* rather than *learning,* to emphasize the relation of behavioral change to biological evolution. Just as organisms evolve in their communities and communities evolve in the wider environment, so patterns of behavior evolve in an individual's lifetime. Evolution occurs on many levels. In biological evolution, replication and variation occur mostly on a genetic level whereas selection acts on individual organisms. In behavioral evolution, replication and variation occur on a biological level whereas selection occurs on a behavioral level; we are born with or develop innate behavioral patterns. But those patterns are shaped by the environment over an organism's lifetime and often attain high degrees of complexity. This is nothing more than a repetition of what every behavior analyst knows. But it is worth emphasizing that operant conditioning is itself an evolutionary process (see Staddon & Simmelhag, 1971, for a detailed empirical study and argument of this point).

Giving Watson Self-Control and Altruism

People often prefer smaller, sooner rewards to larger, later ones. A child, for example, may prefer one candy bar today to two candy bars tomorrow. An adult may prefer $1,000 today to $2,000 5 years from today. In general, the further into the future a reward is, the less it is worth today. A bond that pays you $2,000 in 5 years is worth more today than one that pays you $2,000 in 10 years. The mathematical function that relates the present value of a reward to its delay is called a *delay-discount (DD) function.*

Delay-discount functions can be measured for individuals. Different people discount money more or less steeply than others (they have steeper or more shallow DD functions); these functions can be measured and used to predict degree of self-control. As you would expect, children have steeper DD functions than adults; gamblers have steeper DD functions than nongamblers; alcoholics have steeper DD functions than nonalcoholics; drug addicts have steeper DD functions than nonaddicts; students with bad grades have steeper DD functions than students with good grades, and so forth (Madden & Bickel, 2009). Watson's behavior, to be human behavior, needs to be describable by a DD function too. It would be easy to build such a function (hyperbolic in shape as it is among humans and other animals) into Watson II. But DD functions also change in steepness with amount, quality, and kind of reward. We would have to build in hundreds of such functions, and even then it would be difficult to cover all eventualities. A better way to give Watson II DD functions would be to first give it the ability to learn to pattern its behavior over extended periods of time. As Watson II learned to extend its behavior in time, without making every decision on a case-by-case basis,

it would, by definition, develop better and better self-control. A Watson II with self-control would decide how to allocate its time over the current month or year or 5-year period rather than over the current day or hour or minute. Patterns of allocation evolve in complexity and duration over our lifetimes by an evolutionary process similar to the evolution of complexity (e.g., the eye) over generations of organisms. This ability to learn and to vary temporal patterning would yield the DD functions we observe (Locey & Rachlin, 2011). To have human self-control, Watson II needs this ability.

A person's relation with other people in acts of social cooperation may be seen as an extension of his or her relation with his or her future self in acts of self-control. That there is a relation between human self-control and human altruism has been noted in modern philosophy (Parfit, 1984), economics (Simon, 1995), and psychology (Ainslie, 1992; Rachlin, 2002). Biologists have argued that we humans inherit altruistic tendencies, that evolution acts over groups as well as individuals (Sober & Wilson, 1998). Be that as it may, altruistic behavior may develop within a person's lifetime by behavioral evolution, learning patterns (groups of acts) rather than individual acts. There is no need to build altruism into Watson. If Watson II has the ability to learn to extend its patterns of behavior over time, it may learn to extend those patterns over groups of individuals. The path from swinging a hammer to building a house is no different in principle than the path from building a house to supporting one's family. If Watson can learn self-control, then Watson can learn social cooperation.

Watson in Love

Watson II would not reproduce or have sex.[8] Given his inorganic composition, it would be easier in the foreseeable future for IBM to manufacture his clones than to give him these powers. Would this lack foreclose love for Watson II? It depends what you mean by love. Let us consider Plato on the subject. Plato's dialogue, *The Symposium*, consists mostly of a series of speeches about love. The other participants speak eloquently praising love as a nonmaterial god. But Socrates expands the discussion as follows:

> "Love, that all-beguiling power," includes every kind of longing for happiness and for the good. Yet those of us who are subject to this longing in the various fields of business, athletics, philosophy and so on, are never said to be in love, and are never known as lovers, while the man who devotes himself to what is only one of Love's many activities is given the name that should apply to all the rest as well. (Symposium, 305d)

What Socrates is getting at here, I believe, is the notion, emphasized by the 20th-century Gestalt psychologists, that the whole is greater than the sum of its parts; that combinations of things may be better than the sum of their components. To use a Gestalt example, a melody is not the sum of a series of notes. The pattern of the notes is what counts. The melody is the same and may have its emergent value in one key and another with an entirely different set of notes.

A basketball player may sacrifice his or her own point totals for the sake of the team. All else being equal, a team that plays as a unit will beat one in which each individual player

[8] Robots (or pairs or groups of them) may design other robots. But I am assuming that IBM or a successor will continue to manufacture Watson throughout its development. In the introduction I claimed that Watson has no interests different from IBM's. But that does not mean that Watson could not evolve into Watson II. IBM, a corporation, itself has evolved over time. If Watson's development is successful, it will function within IBM like an organ in an organism, which, as Wilson and Wilson (2008) point out, evolve together at different levels. The function of the organ subserves the function of the organism because if the organism dies the organ dies (unless it is transplanted).

focuses solely on his or her own point totals. Teams that play together, teams on which individuals play altruistically, will thus tend to rise in their leagues. In biological evolution, the inheritance of altruism has been attributed to natural selection on the level of groups of people (Sober & Wilson, 1998; Wilson & Wilson, 2008). In behavioral evolution, Locey and I claim, altruism may be learned by a corresponding group-selection process (Locey & Rachlin, 2011).

According to Plato, two individuals who love each other constitute a kind of team that functions better in this world than they would separately. The actions of the pair approach closer to "the good," in Plato's terms, than the sum of their individual actions. How might Watson II be in love in this sense? Suppose the Chinese built a robot named Mao. Mao, unlike Watson II, looks like a human and moves around. It plays ping-pong, basketball, and soccer and of course swims. It excels at all these sports. It is good at working with human teammates and at intimidating opponents, developing tactics appropriate for each game. However, good as Mao is, wirelessly connecting him to Watson II vastly improves the performance of each. Mao gets Watson II's lightning fast calculating ability and vast memory. Watson II gets Mao's knowledge of human frailty and reading of human behavioral patterns, so necessary for sports, not to mention Mao's ability to get out into the world. Watson II, hooked up to Mao, learns faster; he incorporates Mao's experience with a greater variety of people, places, and things. Watson II is the stay-at-home intellectual; Mao is the get-out-and-socialize extrovert.[9]

It is important that each understand the other's actions. Watson II will know Chinese as well as English.[10] Each will warn the other of dangers. Each will comfort the other for failures. Comfort? Why give or receive comfort? So that they may put mistakes of the past behind them and more quickly attend to present tasks. Mao will fear separation from Watson II and vice versa. Each will be happier (perform better) with the other than when alone. To work harmoniously together, each machine would have to slowly learn to alter its own programs from those appropriate to its single state to those appropriate to the pair. It follows that were they to be suddenly separated, the functioning of both would be impaired. In other words, such a separation would be painful for them both. Watson II, with annoying lights and sounds; Mao, in Chinese, would complain. And they each would be similarly happy if brought together again. Any signs that predicted a more prolonged or permanent separation (e.g., if Mao should hook up with a third computer) would

[9] One is reminded of the detective novelist Rex Stout's pairing of Nero Wolfe (so fat as to be virtually housebound, but brilliant) and Archie Goodwin (without Wolfe's IQ but full of common sense, and mobile). Together they solve crimes. It is clear from the series of novels that their relationship is a kind of love, more meaningful to Archie, the narrator, than his relationships with women.

[10] In an influential article, the philosopher John Searle argued that it would be impossible for a computer to understand Chinese (Searle, 1980). Searle imagines a computer that memorized all possible Chinese sentences and their sequential dependencies and simply responded to each Chinese sentence with another Chinese sentence as a Chinese person would do. Such a computer, Searle argues, would pass the Turing test but would not know Chinese. True enough. Searle's argument is not dissimilar to Dennett's (1978) argument that a computer cannot feel pain. But, like Dennett with pain, Searle ignores the function of knowing Chinese in the world. Contrary to Searle, a computer that could use Chinese in subtle ways to satisfy its short and long-term needs: to call for help in Chinese, to communicate its needs to Chinese speakers, to take quick action in response to warnings in Chinese, to attend conferences conducted in Chinese, to write articles in Chinese that summarized or criticized the papers delivered at those conferences. That computer would know Chinese. That is what it *means* to know Chinese, not to have some particular brain state common among Chinese speakers.

engender still greater pain. For the handlers of both computers, as well as those of the computers themselves, the language of love, jealousy, pain, and pleasure would be useful.

But, as with pain, love *alone* could not make a computer human. Perception, thought, hope fear, pain, pleasure, and all or most of the rest of what makes us human would need to be present in his behavior before Watson II's love could be human love.

What We Talk About When We Talk About Love

Let us consider whether Watson II might ever lie about his love, his pleasure, his pain, or other feelings. To do that, we need to consider talk about feelings separately from having the feelings themselves. In some cases they are not separate. In the case of pain, for example, saying, "Ouch!" is both verbal expression of pain and part of the pain itself. But you might say, "I feel happy," for instance, without that verbal expression being part of your happiness itself. We tend to think that saying, "I feel happy," is simply a report of a private and internal state. But, as Skinner (1957) pointed out, this raises the question of why someone should bother to report to other people his or her private internal state, a state completely inaccessible to them, a state that could have no direct effect on them and no meaning for them. After all, we do not walk down the street saying, "The grass is green" or "The sky is blue," even though that may be entirely the case. If we say those things to another person, we must have a reason for saying them. Why then should we say, "I am happy," or "I am sad," or "I love you"? The primary reason must be to tell other people something about how we will behave in the future. Such information may be highly useful to them; it will help in their dealings with us over that time. And if they are better at dealing with us, we will be better at dealing with them. The function of talking about feelings is to predict our future behavior, to tell other people how we will behave. Another function of Watson II's language may be to guide and organize his own behavior, as we might say to ourselves, "This is a waltz," before stepping onto the dance floor. But how do we ourselves know how we will behave? We know by experience. In this or that situation in the past, we have behaved in this or that way and it has turned out for the good (or for the bad). What may be inaccessible to others at the present moment is not some internal, essentially private, state but our behavior yesterday, the day before, and the day before that. It follows that another person, a person who is close to us and observes our behavior in its environment from the outside (and therefore has a better view of it than we ourselves do), may have a better access to our feelings than we ourselves do. Such a person would be better at predicting our behavior than we ourselves are. "Don't bother your father, he's feeling cranky today," a mother might say to her child. The father might respond, "What are you talking about? I'm in a great mood." But the mother could be right. This kind of intimate familiarity, however, is rare. Mostly we see more of our own overt behavior than others see. We are always around when we are behaving. In that sense, and in that sense only, our feelings are private.

Given this behavioral view of talk about feelings, it might be beneficial to us at times to lie about them. Saying, "I love you," is a notorious example. That expression may function as a promise of a certain kind of future behavior on our part. If I say, "I love you," to someone, I imply that my promised pattern of future behavior will not just cost me nothing but will itself be of high value to me. It may be however, that in the past such behavior has actually been

costly to me. Hence, I may lie. The lie may or may not be harmless but, in either case, I would be lying not about my internal state but about my past and future behavior.

You could be wrong when you say, "I love you," and at the same time not be lying. As discussed previously, our perception (our discrimination) between present and past conditions may lack perspective. (This may especially be the case with soft music playing and another person in our arms.) Thus, "I love you" may be perfectly sincere but wrong. In such a case you might be thought of as lying to yourself. The issue, like all issues about false mental states, is not discrepancy between inner and outer but discrepancy between the short term and the long term.

So, will Watson II be capable of lying about his feelings and lying to himself about them? Why not? Watson II will need to make predictions about his own future behavior; it may be to his immediate advantage to predict falsely. Therefore, he may learn to lie about his love just as he may learn to lie about his pain as well as his other mental states. Moreover, it will be more difficult for him to make complex predictions under current conditions when time is short than to make them at his leisure. That is what it takes to lie to himself about his feelings.

DOES THE MECHANISM MATTER?

Let us relax our self-imposed restraint of appearance and bodily movement and imagine that robotics and miniaturization have come so far that Watson II (like Mao) can be squeezed into a human-sized body that can move like a human. Instead of the nest of organic neural connections that constitutes the human brain, Watson II has a nest of silicon wires and chips. Now suppose that the silicon-controlled behavior is indistinguishable to an observer from the behavior controlled by the nest of nerves. The same tears (though of different chemical composition), the same pleas for mercy, the same screams of agony as humans have are added to the behavioral patterns discussed previously. Would you say that the nest of nerves is really in pain but the nest of silicon is not? Can we say that the writhing, crying man is really in pain but the similarly writhing, crying robot is not really in pain? I would say no. I believe that a comprehensive behavioral psychology would not be possible if the answer were yes; our minds and souls would be inaccessible to others, prisoners within our bodies, isolated from the world by a nest of nerves.

Nevertheless, many psychologists and many behaviorists will disagree. (They would be made uncomfortable with Dolly II, probably get a divorce, were she to reveal to them, perfect as she was, that she was manufactured, not born.) Why would such an attitude persist so strongly? One reason may be found in teleological behaviorism itself. I have argued that our status as rational human beings depends on the temporal extent of our behavioral patterns. The extent of those patterns may be expanded to events prior to birth, to our origins in the actions of human parents, compared to Watson II's origins in the actions of IBM. Those who would see Watson II as nonhuman because he was manufactured, not born, might go on to say that it would be worse for humanity were we all to be made as Watson II may be made. To me, this would be a step too far. We are all a conglomeration of built-in and environmentally modified mechanisms anyway. And no one can deny that there are flaws in our current construction.

REFERENCES

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. New York: Cambridge University Press.

Baum, W. M. (2005). *Understanding behaviorism: Science, behavior and culture*. New York: Harper Collins.
Block, N. (1981). Psychologism and behaviorism. *Philosophical Review*, *90*, 5–43.
Dennett, D. (1978). Why you can't make a computer that feels pain. *Synthese*, *38*, 415–416.
Grunow, A., & Neuringer, A. (2002). Learning to vary and varying to learn. *Psychonomic Bulletin and Review*, *9*, 250–258.
Locey, M. L., & Rachlin, H. (2011). A behavioral analysis of altruism. *Behavioural Processes*, *87*, 25–33. NIHMS 260711.
Madden, G. J., & Bickel, W. K. (Eds.). (2009). *Impulsivity: Theory, science, and neuroscience of discounting*. Washington DC: American Psychological Association.
Moyer, M. (2011). Watson looks for work. *Scientific American*, *304*, 19.
Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*, 435–450.
Noë, A. (2009). *Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness*. New York: Hill and Wang.
Parfit, D. (1984). *Reasons and persons*. Oxford University Press, Oxford.
Rachlin, H. (1985). Pain and behavior. *Behavioral and Brain Sciences*, *8*, 43–52.
Rachlin, H. (1994). *Behavior and mind: The roots of modern psychology*. New York: Oxford University Press.
Rachlin, H. (1995). The value of temporal patterns in behavior. *Current Directions*, *4*, 188–191.
Rachlin, H. (2000). *The science of self-control*. Cambridge, MA: Harvard University Press.
Rachlin, H. (2002). Altruism and selfishness. *Behavioral and Brain Sciences*, *25*, 239–296.
Ramachandran, V. S. (2011). *The tell-tale brain: A neuroscientist's quest for what makes us human*. New York: Norton.
Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*, 417–424.
Simon, J. (1995). Interpersonal allocation continuous with intertemporal allocation. *Rationality and Society*, *7*, 367–392.
Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.
Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.
Staddon, J. E. R., & Simmelhag, V. L. (1971). The "superstition" experiment: A reexamination of its implications for the principles of adaptive behavior. *Psychological Review*, *78*, 3–43.
Tononi, G., & Edelman, G. M. (1998). Consciousness and complexity. *Science*, *282*, 1846–1851.
Wilson, D. S., & Wilson, E. O. (2008). Evolution "for the good of the group." *American Scientist*, *96*, 380–389.