The imperfect observer: Mind, machines and materialism in the 21st century

Judith Donath
MIT Media Lab
judith@media.mit.edu

1. Introduction

I was asked to write this paper on dualism from the perspective of the "natural sciences"[1]. Most contemporary scientists are physicalists, believing that the universe is entirely physical, and it is from this perspective that I am writing[2]. I am going to focus on one of the most contested areas in the dualist/physicalist debate: the nature of mind.

When asked to write such a paper, one of the first questions that comes to mind is "Why me?" What can I add to the immense literature on this subject? What fresh and useful perspective can I bring?

I am designer and theorist working in the field of computer mediated communication. I am interested in the complexity of human communication, in questions such as how do we maintain truthfulness when deception can be so profitable? and How do we form impressions of each others identity? One theme that runs through this work is the imperfectness and subjectivity of our perception. It is this concept that I am exploring in this paper.

---

[1] This paper was prepared for a colloquium on dualism, part of a three part series on "Questioning 19th-Century Assumptions about Knowledge", sponsored by the Gulbenkian Foundation and organized by Immanuel Wallerstein (Yale University), Aviv Bergman (Albert Einstein College of Medicine) and Jean-Pierre Dupuy (École Polytechnique, Paris).

[2] I believe that the mind is fully a function of the embodied brain – that it, the brain and its context within the body and the sensory nervous system. We grow from single cells – one can look at the blastula as it grows, slowly developing a nervous system. While I do not believe that this early embryo "thinks" and I do believe that babies do, I do not think there is a magic switch that occurs, but rather a slow continuum of increasing sensory capability and awareness, combined with a growing memory. I believe that physical events in the brain – neural and chemical -- cause all sensation we have. I think we are far from understanding the complexities of this process, and there are vast open questions about how memories are formed, how sensory data is translated into impressions of the world etc. And I am not sure whether we can ever fully understand the relationship between subjective experience and observed brain – I think we are perhaps missing the right way of looking at the problem.

Physicalism[3] states that the world, including the mind, is entirely physical, and that, in theory, it should be entirely knowable through non-metaphysical methods. But human beings are not perfect observers and our ways of perceiving and interpreting the world have inherent limitations. And perceiving the extent of these limitations is one of the things we may not be able to do very well.

In this paper I am looking at the observer – at the question of, in a physicalist view of the world, how capable are we of really understanding that world. In traditional western dualistic, there is less emphasis on observation because much of what is of greatest interest – the mind, the spiritual, god, etc. – is assumed not to be observable, either inherently unknowable or understandable only through revelation or intuition. The physicalist view places great emphasis on observation; in the scientific framework a world that is wholly physical should be ultimately fully observable. Of course, that view has been challenged, most notably in physics, where quantum indeterminacy says that at the quantum scale a system is inherently indeterminate and not completely measurable; the uncertainty is inherent in the physics of the system. Here, I am focusing on something else – not whether the external world (at the classical rather than quantum level) is knowable, but whether it is knowable *by us*; that is, what are the limitations of observation and comprehension that are inherent to being human – in particular, in the case of trying to understand the questions of mind and consciousness in a physicalist framework

This paper is in three sections. I first outline some of the problems in the dualist/physicalist debate about the nature of mind, next I discuss some findings from neuroscience and the insights they provide about our subjective perception, and finally I look at the quest to build intelligent (seeming) machines and its philosophical and practical implications regarding the knowability of minds.

2. Dualism, physicalism and the limits of perception

Summary: Beyond the argument about the actual nature of the universe (and/or the mind) is further debate about what are the value and the limits of observation. One can be a physicalist and still believe that not everything is humanly observable and deducible: the world may be entirely material, yet our perceptual and/or cognitive systems are too limited to fully understand it, indeed perhaps at best allowing only a tiny and distorted glimpse of it.

Outside of the mind, there is the world of external reality. It may be a wholly material world, propelled and bounded by physical laws, or it may be a world of both material and spiritual essences, propelled by the hands of God or gods, but whatever form it has, it constitutes an external and unshakeable reality, one that people may observe and conjecture about, but the shape of their beliefs does not shape the basic structure of the universe[4]. Rather, these beliefs may be right or wrong, depending on how accurately they model the external reality, though owing to the limitations of human observation, this may not be possible to settle decisively.

[3] I am using the phrase physicalism rather than the somewhat more common materialism to avoid confusion with other meanings of materialism (e.g. rampant consumerism).

[4] This in itself is a contestable statement. Extreme skeptics doubt the possibility of observation, so trapped we are within the limits of our mind; quantum influenced philosophy emphasizes the transformative and distorting power of the observer's gaze.

The limitations of human observation are both perceptual and cultural. Perceptually, we are capable of observing a limited band of phenomenon: things that can be heard (sonic vibration in the frequency range of 20 to 20,000 Hz), things that can be seen (light waves in the frequencies of about 380 to 740 nanometers), felt, smelt or tasted (and, perhaps, intuited, depending on one's belief in both the existence of non-material phenomena and, if that exists, in the human ability to in some way perceive it). Even at the level of basic perception it must be noted that what we see is a cognitive interpretation of the external: a particular wavelength of light provides us with a sensation of seeing the color green, but the experience of that physical phenomenon as color is a feature of our perceptual system, not of the light. Other species and indeed other individuals may see it differently; presumably, other creatures could experience that wavelength of energy in a completely different modality, rendered like our experience of taste or sound or something unimaginable to us. Beyond this perceptual subjectivity is cultural subjectivity: the worldview in which we were raised and the beliefs of the people around us profoundly affects our interpretation of what we experience.

This state of affairs leaves people with a broad menu of possible beliefs. One may believe that the universe is purely material or that it is dualistically both material and spiritual; in either case one may believe that humans are capable of fully understanding it or one may believe that vast parts of it are forever outside of human comprehension.

These issues come together in the debate about the nature of mind. Is the mind, and our sensation of consciousness, an effect purely of the physical brain and its material context, or is there some non-physical aspect to consciousness? Descartes most clearly set forth the position that the mind, i.e. consciousness and the self, is a separate substance than the brain, existing outside of the physical world; this is the foundation of dualism in philosophy of mind. The opposing, monistic, argument is that everything is physical, including thus any states, sensations or substances ascribed to mind or consciousness.

Most people, other than scientists and some philosophers, are dualists. It seems intuitively true that there is some non-material, conscious core to one's being, separate from the body (Bloom 2004). Some, especially those whose perspective is religious, believe that the immaterial self is the soul, which may live on beyond the body; this spiritual dualism generally exists within a fully dualistic context in which the universe consists of both God and matter. Yet dualists need not be religious: one may believe in an immaterial self without believing in God or in an eternal soul. We are conscious, self-conscious beings and the experience of being human gives rise to a strong sense of a distinct mental self (d'Andrade 1987).

Today, most scientists and many philosophers are physicalists. If one believes that the universe is composed entirely of matter, then it seems intuitive that the mind is also a fully physical entity. If not, how would this non-material entity arise within the course of evolutionary development – where in the chain of events leading from the earliest carbon-based life forms to sponges and flatworms and on to cynodonts and the early mammals (and the development of the neo-cortex) leading to the primates and eventually to humans – where did some super-material conscious entity arise and how?

Much of the debate centers on the significance of one's own experience of consciousness. I perceive sensations such as pain or the color blue or my memory of last night's dinner from a fundamentally personal and experiential viewpoint. Today, few would argue that these sensations have not occurred because of something happening in the brain, something physical. What is debated is the meaning of the personal experience: is it an emergent property of the complex physicality of the brain, different perhaps in

scale than the reflexes of the lobster, but not an essentially different phenomenon, or is it something that occurs in a fundamentally different plane, one that is separate from, though intertwined with, the physical?

If one believes the dualist premise that the mind is separate and non-material, then one generally also believes that external observation of physical states is not capable of perceiving the mind. However, this does not mean that mind is entirely unobservable: one's own mind, one's own state of consciousness, experience of pain, etc. are quite accessible through subjective feeling. Contemporary philosophy has termed these experiences "qualia" and for the dualists, they are evidence that the mind exists and is of a different type than the brain. For the dualist, the personal, subjective experience of internal mental state is significant and the external, purportedly objective observation of the brain may be instructive, but is ultimately limited in what it can reveal about the mind.

If one believes the physicalist premise that the mind, that is thoughts, feelings, cognition, the sense of self, arises from and is wholly caused by the brain, then it follows that one believes that any change in "mind" – i.e., any new perception, emotion, memory access, etc – originates in a physical change in the brain. Given sufficiently good observational tools we should be able to see the corresponding physical events for any change in mental state. Here, it is the personal, subjective experience of internal mental state seems to be limited and possibly misleading, while the external, objective observation of that brain promises to yield a deep understanding of the nature of consciousness.

At first glance, the physicalist approach to mind would seem to be simple: the brain is the cause of all thought, including the sense of consciousness, and while we may not have the tools today to observe all the brain's functions, they are not inherently unobservable and as better tools become available, eventually we will understand all.

It is, however, more complex. Although the physicalists agree that consciousness is not a separate entity (James 1904), what exactly it is debatable. Some believe that consciousness is simply an effect of the brain's activity and what we think of as consciousness is an illusory experience (Dennett 1988). Others see consciousness as a significant, but inherently private understanding, one that can be observed but the actual experience of which cannot be shared or directly observed (Crick and Koch 1998). Others believe that with the right tools, models and theories, the relationship between the activity in the brain and the experience of being will be fully elucidated (Ramachandran and Hirstein 1998).

As neuroscience reveals more about the brain, it is possible that we will come to a better understanding of the limits of our cognition, as well as gaining insight about how the structure and activity of the brain give rise to the experience of consciousness (Crick and Koch 1998). For instance, there are numerous neuroscience studies that are examining the neural correlates to conscious experience (e.g. (Hohwy and Frith 2004) and through these we may learn also new intuitions about how thought works, including the difficult problems about what it means for consciousness to be private.

Whether discovering the neural correlates of consciousness answers the "hard problem" (Chalmers 1995) of consciousness is the subject of intense debate. For some, these correlates ARE the answer to the question of what is consciousness. Others, who grant that the brain states are necessary for the conscious state, remain unconvinced that the observation and description of neural behavior is sufficient for explaining consciousness. Some feel that the two are at different levels of description; others feel that experience is simply private and will have to remain that way. Finally, there are dualists, such as Chalmers, who propose that consciousness is irreducible to physical state but that it comprises or is composed of an additional fundamental natural type, which he suggests may be information.

2.1. Mary sees red

One well-known thought experiment that argues against physicalism is Frank Jackson's story of Mary, a hypothetical highly educated scientist of the future who has lived all her life in a black and white room[5]. Mary lives at time when everything about the physical world, from physics to neurophysiology, is known, and she learns it all, via black and white media. She knows everything about how everything works, but she has never seen the color red. And then one day she is introduced to world of color. Jackson's claim is[6] that she will learn something from this experience – she will learn what it is like to experience qualia such as the colors red, purple, blue etc. – and therefore physicalism is wrong (Jackson 1986). Jackson's intent was to argue against reductionist physicalism, to show that we cannot derive phenomenological experience from physical knowledge and therefore such experience must be extra-physical.

This parable has spawned a number of responses (many collected in Ludlow, Nagasawa, and Stoljar 2004). Daniel Dennett, for instance, argued that if Mary truly knew *everything* about color, that knowledge would include an understanding so deep that it would allow comprehending qualia from the knowledge about them; we are so far from having such knowledge that we do not have the intuition for understanding its implications.

Dennett's goal was to defend physicalism. I would like to take a somewhat different angle here, and use this thought experiment to highlight the difference between what the world is and how the world is perceived. Physicalism is addresses the question of what the world is: it says that everything is physical and no substances exist outside the physical realm. However, it does not guarantee that we as observers are capable of perceiving or understanding everything about that world. I think the fallacy in the thought experiment is that it leads us to think of "all knowledge" as a massively extended encyclopedia, a giant compendium of the type of knowledge that we as contemporary human beings have. We know that our current knowledge is limited - we're still discovering new stars, and are in the midst of decoding the genome, and are just beginning to map the brain and its complex networks. So there is a lot more knowledge, of the type that we have currently, to still be discovered. And perhaps somewhere in the extension of that knowledge will be the tools and techniques that, as Dennett said, would allow one to imagine a hitherto unseen "quale".

But it is also possible that there are inherent limitations in our ability to make sense of the world – that we can learn more and more, develop all kinds of tools and techniques, and yet be far from understanding the universe, or even our own minds, in their totality. Such a limitation may be unsurpassable, but it is a function of the observer's ability, not the nature of the universe.

---

[5] This thought experiment, along with Nagel's "What is it like to be a bat?"(Nagel 1974) are often referred to as the Knowledge Argument, because they claim that not all knowledge is reducible to physical facts.

[6] Actually, Jackson's claim *was.* He has since revised his views and is a proponent of physicalism.

Indeed the Mary the color deprived neuroscientist experiment highlights how impoverished and logo-centric our conception of "knowledge" is. Let us keep the experiment as is, with its black and white walls, and books and TV. But let us also say that at this point in the future a lot of our communication occurs via direct brain stimulation. Since we have learned so much about the brain, we can set up an electrical field helmet (or whatever device would be appropriate) and create brain states that mimic experience, in this case, the experience of seeing red. So while Mary has never seen red visual, she has had the experience of it (much as you might see something as bright red in a dream, though you are in fact in a dark room with your eyes closed)[7]. One might argue that such neural stimulation is not a legitimate mode of knowledge acquisition; such arguments help to highlight that the claim of all-inclusive knowledge is in fact a limited "description of" the world. It forces us to examine why we perceive that a textbook conveys knowledge and an electric probe does not – particularly if one could have the same experience from reading the book or undergoing the stimulation. In any case, the fact that such a description is incomplete – that it does not handle qualia (and probably many other phenomenon that occur outside of our perception) does not mean that the world is not entirely physical. It means the world might not be encompassed within a particular type of knowledge collection – that of conscious, communicable, human cognition.

2.2. Can neuroscience see the mind?

For the physicalist, there is nothing inherently unknowable about the world: all states and properties can, somehow, be perceived. The limits of the unaided human eye are well known, and thus we make tools that extend our observational reach: from high-energy particle telescopes to scanning electron microscopes. And we are now making tools for seeing the brain at work - CAT scans, PET scans, functional magnetic resonance imaging – techniques that let us observe the process of thought.

But are we truly capable of seeing thought? Even with the assumption that there is no super-physical aspect to it, are we as observers capable of seeing it? First, there is the physical difficulty: neuro-imaging, while advancing at a fantastic rate, is still quite primitive. For example, FMRI, one of the most advanced current techniques, measures blood flow to areas of the brain that have high neural activity. Yet neural activity is the real interest and the blood flow, while non-invasively detectable with relatively good spatial resolution, is an indirect and slow indicator of where that activity occurs. Presumably, imaging techniques will continue to improve and decades from now, scientists (as well as employers, police officers, suspicious spouses, etc.) will be able quickly, painlessly, and accurately to see the electrical and chemical activity in other people's brains.

Neuroscientists map the brain by giving subjects a task – raise your right arm, watch this violent movie, read this story and tell us if anyone lied in it – and noting the areas of with high activity. This and

---

[7] Churchland points out in his argument against the Knowledge Argument that someone who has not seen color their entire life would be unable to see it when they left the b&w room as the neural synapses for perceiving color would not have developed (Churchland 1989). I'll assume for the sake of this argument that she is able to perceive the induced color – we can imagine, for instance, that she was stimulated with this "fake" color experience from an early age and thus developed normally….

other techniques that reveal what parts of the brain perform which functions provide the neural correlates to thought and action, but alone this is arguably not tantamount to understanding thought.

Yet mapping the brain does provide the foundation for understanding how primitive responses and affective processes contribute to complex conscious experience. Seeing that an area of the brain known for disgust or pleasure plays a role in social interaction and motivations sheds light on how these behaviors and abilities have developed.

For example, researchers have looked at brain activity during rounds of the classic experimental economics task, the ultimatum game. Here, two subjects have a single interaction. The first is given a sum of money and told to divide it with the second. If the second accepts the offer, they each get to keep the divided amount; if the second rejects the offer, they both get nothing. Generally, the results are that if the division is close to equal, the offer is accepted and both parties get something, but if the division is too uneven, the offer is rejected. In terms of traditional economics, this provided evidence that people behave irrationally: would it not be better to get, say, $1 instead of nothing, regardless of what one's partner got? Remember, the experiment is a one-time interaction, so the rejection is not a bargaining strategy. Understanding this behavior requires putting it in a social context, where factors such as anger and the desire to punish an unfair partner, vie with the desire for economic gain in how subjects make their decision. Yet the results are still a bit puzzling: why would someone take a personal loss (of the potential earning from the game) in order to punish someone they do not know and will not interact with again?

Seeing what is occurring in the brain during rounds of the ultimatum game provides some answers (Sanfey et al. 2003). When given a low offer, people showed activation in the anterior insula, a part of the brain that activates during times of pain and distress; indeed, the neural response to a low offer was quite similar to the response prompted by encountering a disgusting object. The stronger the response in this area, the more likely it was that they would reject the offer.

One interesting aspect of such studies is that they show how the higher level functioning of the brain builds upon lower level structures and functions. Here the foundation of the emotional response to an upsetting social experience is the earlier brain structures that produce the emotional response of disgust and anger.

The anterior insula is part of the cerebral cortex, but is a relatively older part of this high-level brain structure. It is plays a role in producing an emotional response to sensory stimuli. Animals with limited or no cerebral development, such as fish, can respond to pain or other sensory input, but, it is believed, do not have a conscious or an emotional experience of it (Rose 2002). The development of the cortex in mammals (and the cortical regions of birds and some other vertebrates) made emotional responses to such stimuli possible, such as disgust at the smell of rotting food or fear at the sight of a predator. Indeed, having an emotional response to stimuli, rather than simply a behavioral reaction, is one of the foundations of consciousness.

Neural imaging studies also show how the brain influences the decisions and behavior that lie at the heart of human sociability and of the "higher aspects" of thought. In 2004 Quervain et al (Quervain et al. 2004) published a paper entitled "The Neural Basis of Altruistic Punishment". Altruism means acting for the benefit of others at the cost to oneself, and many religions hold being altruistic to be a defining characteristic of leading a good life. Altruism also poses a puzzle for evolutionary biologists, economists, and others who predicate their models of behavior on individuals always acting to maximize personal

benefits and minimize costs. Given such a model, altruism simply should not exist. Yet seemingly altruistic behaviors do occur, and they are not peripheral oddities, but are at the core of social cooperation.

For example, people altruistically punish violators of social norms. If we look only at external costs and benefits it appears that those doing the punishing take on the cost of doing so without any personal gain: others will primarily benefit from future interactions with the now chastened and presumably better-behaved miscreant. Quervain and her colleagues have found that a part of the brain (the caudate structure of the anterior dorsal striatum) associated with decisions with anticipated reward is activated upon enacting punishment on a social defector. They found that subjects with higher caudate activation were willing to spend more resources to punish defectors and hypothesized that subjects anticipated significant psychological rewards.

Quervain et al also noted that this experiment highlights the difference between the biological and psychological (though perhaps this might be better termed philosophical) definition of altruism. In biology, an altruistic act is one that benefits others while being costly for the actor; it does not involve the intention of the actor. For some philosophers, however, in order to qualify as altruistic, an act must be motivated without expectation of reward, whether external or internal. While such experiments do not rule out the existence of strictly altruistic acts, they do show that rewards generated within the brain motivate some, if not all, altruistic acts.

2.3. The metaphorical brain

In 1980 the linguist George Lakoff with his colleague Johnson (Lakoff and Johnson 1980) proposed that all thought is grounded in physical experience and the more abstract ways we have of thinking are based on and constrained by our physical experience. According to this model, we build our complex intellectual edifice from a foundation in immediate perception. For example, we are "up" when happy and "down" when depressed, work can be "going" well or it may have "stalled".

As we learn more about the brain, it appears that the neural processes of experience are in fact analogously metaphorical. For example, we speak of the "pain" of grief, of the "hurt" feelings that come from being excluded. It turns out that this pain is not just a figure of speech, but a sensation that occurs within the brain (Panksepp 2003): people who experience social rejection respond with activity in the anterior cingulated cortex, an area of the brain associated with aversion to physical pain.

A highly simplified model of how the brain has evolved is that the earliest parts of the brain, such as the brain stem, handle the basic essential functions such as respiration and the reflexive response to stimuli. As other parts of the brain evolved, such as the various parts of the brain that provide an emotional response to perceived stimuli, such as pain in response to adverse stimuli, they often took as their input not direct stimuli from the outside world, but activation and other status information from other parts of the brain. Thus, the ability to learn to avoid pain is built on having the affective response of pain, which is in turn built on the low-level perception. Some of the more recently evolved areas of the brain are those that respond to social situations. Here again the brain works within the context of its existing structures and functions. The earlier areas that could generate the sensation of wanting to avoid something (fear, disgust) or of wanting to continue and repeat something (pleasure) become part of subsequent activation networks involving situations such as punishing transgressors (which is pleasurable (Quervain et al. 2004)), being treated unfairly (which is akin to disgust (Sanfey et al. 2003)) and being rejected (which is painful (Panksepp 2003)).

The big question in the context of the physicalist debate is whether these descriptions of how the brain functions answer the questions of what is consciousness, how does it arise in the brain and how are brain states related to experiential states. For those who feel that these studies are indeed enlightening and that they are useful in developing an intuitive link between mind and brain, the field is progressing rapidly. Of those for whom these studies are in some way beside the point - "No one has produced any plausible explanation as to how the experience of the redness of red could arise from the actions of the brain." (Crick and Koch 2003) - the solution is to either assume, as do Crick and Koch, that experience is, while physical, inherently private, or to attribute this "explanatory gap" (Chalmers 1995) to an extra-physical phenomenon.

2.4. What is it like to be you?

At the center is of this issue is the problem of communication. I can observe my own subjective experience quite nicely; the problem comes when I try to convey it to you. If I say "the house is red" we assume that the red you imagine is fairly close to the one I have in mind and we can get closer with comparisons, like "it is the same red as Sam's car". We are assuming that we each experience red in pretty much the same way, but we have no direct way of knowing for sure. The light waves that bounced off Sam's car made a particular impression on me and some other, possibly very similar, impression on you. And we now both know that the house will give each of us a subjective red experience pretty much like the car does – but what we don't know for sure is if each of our subjective experiences matches the others. We assume it is, but we do not know for sure.

Some say that these qualia, these personal experiences, are inherently private (Crick and Koch 1998). Others suggest that advances in science may provide new communication channels that will make it possible to share experience more directly. Ramachandran and Hirstein propose directly connecting neurons from one brain to another, so that one could actually have the experience of being another (Ramachandran and Hirstein 1998). Their argument is that any other form of communicating brain state, whether spoken language or imaging technology output, is a translation and it is in the process of translation that causes the barrier in understanding.

Ramachandran and Hirstein's thought experiment is intriguing, though not entirely convincing. The complex interconnection of the brain's neurons are created over time in the process of building memories and associations, meaning that individuals bring a somewhat different set of connections to each experience. If the receiving brain is different from the brain providing the input, there is still a significant subjective difference in the experience.

However seeing the problem of the imperfect observer as a problem of communication and of translation is very useful. While perfect comprehension of what it is like to be another person may not be possible, it may be feasible to create communication channels that are far more direct than our existing languages[8]. Such research is today quite speculative.

---

[8] Communication between people is indeed a process of translation and reinterpretation, one that is far from the conduit-like imagery Ramachandran's neural cable implies. Yet that is not necessarily a failing: we need to be careful of positing an uninhibited flow of information from one brain to another as

One promising research area is work in mirror neurons (Rizzolatt and Craighero 2004), which are believed to underlie the imitative abilities of humans and other primates. These and other areas of the brain are being studied to understand the neural correlates of empathy (Carr et al. 2003). Whether or not this research will lead to new channels of direct communication is unclear, but it is certainly helping to pinpoint the cues we read in others in order empathize with them – in other words, to create a simulated version of their experience in our own minds.

2.5. Seeing how unclearly we see

The brain imaging work also highlights another question about the reliability of the observer, showing how deeply, subjectively, and peculiarly embodied our entire perceptual system is. While it feels to us that we simply see and understand the outside would as a direct conduit of information, it is actually heavily filtered through the complexity of our cognitive system.

The more we know about how the brain works the more indirect we see our perception of the world to be. We see and comprehend things as patterns, as meaningful events, because of the way our brains operate and they are made for the interpretation of a certain type of information, a certain type of world. When we react with anger and disgust to a low ball offer, those are metaphorical emotions. The complex interconnected activity of the brain creates them, using a sensation that evolved for one purpose to help make sense of a different input that, presumably, acquired some benefit from producing a similar experience. So we start to see the building blocks of our mental experience. But we also see that at an inescapable level, we perceive particular types of patterns – and not others.

2.6. Ethics in the brain

Where do our ethics come from?

A literal answer is that the STS (superior temporal sulcus) region is initial site of social perception. It is here that we process visual information about the actions of other people, such as their gestures, gaze direction and facial expression. This region in turn stimulates others that produce more subtle analyses of social interaction, such as the amygdala, which is involved with handling social emotions (developing and expressing conditioned fear, analyzing facial expression and other emotionally significant social stimuli), and the orbitofrontal cortex (Rolls 2004), which is involved in more abstract social thinking (planning future activity, responding to punishment, social manners, concern and empathy for others) (Allison 2001). Damage to these parts of the brain affects social reasoning and learning, the basis of ethical behavior.

One of the deepest concerns about a physicalist worldview is that it eliminates the basis for ethics. Dualism posits a higher power, an eternal soul. The “death of God” argument is that ethics has suffered a serious blow with the decline of religion, for if we once believed that our ethics were divinely granted, once

---

the ideal for communication. Human communication, whether through language, gesture or display, is far more ambiguous and manipulable - possibly for good reason.

we come to believe that all is physical, that there is no spiritual essence, then the reasons for following ethical precepts disappear along with religion.

One response is a deterministic evolutionary morality (e.g. (Miller 2000) or even (Kropotkin 1902)). We do not ascribe the socially cooperative behavior of a pride of lions or a herd of gazelles to moral considerations; we assume that they have evolved the instincts for this behavior. Similarly, we can see human cooperation as having evolved as a series of biological responses that improved group fitness through cooperative measures, social sanctioning of inappropriate behavior, etc. Recent neuroscience results such as Quervain et al's discovery of the pleasurable component in of altruistic punishment have bolstered this argument by showing that there are reward systems in the individual that help start and sustain cooperative societies.

Yet there is vast empirical evidence that human beings, whether left to their own devices or under the influence of organized religions, whether deeply dualistic and heeding the word of God or resolutely physicalist, very often act in ways quite antithetical to most accounts of ethics. And even in areas that would arguably be among the most biologically determined, such as the care of infants, there is a tremendous range of behaviors, attributable to both individual and cultural differences (Hauser 2006; Hrdy 1999).

Neuroscience can affirm that we become angry when cheated, and it can provide a narrative showing how more primitive emotional responses provide the foundation for modern human's complex social interactions. It is important to keep in mind that seeing such responses is not tantamount to knowing why we have them: they may be deterministically inborn but it can also be that they are the product of freely willed intent. (Brown 2002) argues that neuroscience supports a non-deterministic view of human morality, citing for example research that shows that deliberate behavioral changes in treating obsessive-compulsive disorder had the same eventual effect on neural response as did a regimen of psychoactive drugs. The responses we see in the brain are the result of both the genetically encoded physical structure and its subsequently modified form, reshaped through culture and deliberate action.

The big question is what we do once we can observe the neural foundations of ethical behavior. Do we strengthen them? Overcome them? How will we behave in an era that has the possibility of unprecedented mind control?

2.7. Like our mind, but different

One of the key things to keep in mind is that the difficulty of explaining consciousness through our current means of studying the world should alert us to the possible existence of innumerable phenomena that do not readily make their existence known via contemporary observation. Studying human consciousness is certainly a hard problem, but also one that we are uniquely situated to observe. We bring our own experience of feeling pain, seeing green, being happy to our study of the brain. Even if we cannot agree on the fundamental meaning of the correlation between an observed neural state and subjective experience, at least we know of and can comprehend the subjective experience. Yet many other systems may exist which have some significant aspect that does not reveal itself to human observation.

Such speculation can easily veer into science fiction and ungrounded metaphysical musings about cosmic consciousness and the psychic life of trees. However, it does directly relate to the question of machine consciousness, which I will be addressing in the next section of this paper. As we learn more about the complexity of the brain and it's intricately interconnected neural networks and hormonal baths,

the notion that a machine could think seems silly, for the most complex computer is, comparatively, extremely simple. But their complexity is growing and there is again no inherent reason why a machine cannot one day achieve organic scale complexity. Surrounding this possibility are several questions, some often asked (though more in science fiction than in philosophy): Could it be conscious? How would we know that is was? Might not it have some private state that was completely alien to us – and would we have any access to this private state, even if just by analogy? And the question of machine minds is one of immediate practical importance, for we are living in a world where interactions with machines that act sentient are becoming an everyday event.

3. The intelligent machine

In 1950 Alan Turing published an extraordinarily influential and provocative paper, "Computing Machinery and Intelligence". The paper begins with the statement "I propose to consider the question, 'Can machines think?'", a question that he quickly discards as "too meaningless to deserve discussion", substituting instead a behaviorist approach to the problem: can machines be made to act indistinguishably from humans?

Turing proposed a version of a parlor game called "The Imitation Game" as the venue for judging the machine's behavior. In the original version, a player, who may be male or female, attempts to convince another player, the judge, that he or she is a woman. The players are in separate rooms, communicating only via written messages. In Turing's version, it is a person or a machine that is the hidden player, and whose task it is to convince the just that he/she or it is human.

The paper launched the field of artificial intelligence. Creating a machine that cannot be distinguished from a human turned out to be harder than Turing predicted; he thought it would be done within 50 years, a milestone that has already passed, and we are still quite far from having such an a device. There is a yearly contest, the Loebner Prize, which runs a limited version of the Turing Test (as this game has come to be called), and each year, some of the machines fool some of the judges, but none have yet been very convincing. Yet we are already living in a time when the line between human and machine is blurring. We call a company, seeking information, and the helpful sounding operator at the other end is actually a text-to-speech synthesizer with voice recognition capability. We log into a chat room, and the witty jokester we have been bantering with turns out to be a software agent, combining some lexical analysis with a dictionary of humor. Turing's paper revitalized the ancient question about whether machines could think at a moment in history when technology, in the form of computers, promised to make such an invention possible.

The paper also launched debate about whether the Imitation Game was a good substitute for the question of whether machines can think. To accept it, you must agree with the pragmatic, behaviorist approach: "thinking" is an internal state that is forever private and invisible, thus it is pointless to attempt to fathom it. We can only go by the externally observable behavior. Then, you must agree with Turing's assumption that the use of language in a conversational setting was the aspect of human behavior most relevant to thinking; indeed, that it was sufficient to imply thinking. And finally, you have to believe that the judge's ability to determine which was human was adequately reliable.

The significance of Turing's paper is not because of the worthiness of the game itself, but because each of the points upon which one might argue about it – the behaviorist approach, the importance of language, and the reliability of the judge – are significant both philosophically and practically.

Turing's behaviorism was not cynical. The machines he envisioned playing the Imitation Game were complex and sophisticated, complete with a vast compendium of knowledge and good reasoning power. His agnosticism about their thinking ability was about whether the process they used to enact their role was analogous to ours; he seems to believed that it might be quite different, but on par in terms of complexity.

In 1966, Joseph Weizenbaum, a computer scientist at MIT, created ELIZA, the first conversational computer program. Unlike Turing's envisioned thinking machine, ELIZA was a simple text processor, programmed with rules for grammatically transforming sentences and a set of known keywords that could trigger special responses. It also was equipped with a new persona, that of a Rogerian psychiatrist.

Weizenbaum's goal was most certainly not to propose that the future of psychoanalysis lay with machine therapists. Rather, he wanted to demonstrate that a machine could engage in human-like interactions without its underlying program having any pretense of intelligence (Weizenbaum 1966); it was to be a refutation of Turing's contention that human-like language based interaction was in any way a measure of intelligence. By suggesting the Rogerian psychiatrist as the model for ELIZA, he created a context in which the program's limited knowledge and conversational ability made sense.

ELIZA was both a great success and a great failure. People enjoyed conversing with her/it, and even those who knew the technology behind it became caught up in discussing their problems with the "psychiatrist program". Thus, it succeeded at showing that a heuristically based parser program could indeed be an engaging conversationalist. Yet people went further, seeing it as a useful therapeutic tool and even predicting that programs like it would replace human psychiatrists. This horrified Weizenbaum, who had intended the opposite response: that people would understand that their own projection was the source of any intelligence they perceived in the machine and that they would thus reject it and other such programs for any task requiring intelligence, let alone empathy.

The case of ELIZA highlights the problem of the imperfect observer. It is a type of behavior quite tractable to programmatic modeling: the program need not have an immense knowledge base; the ability to parse and rephrase sentences, plus some keyword-based heuristics will do. Yet by framing ELIZA as a Rogerian psychiatrist, Weizenbaum created a situation in which people's expectations of ELIZA's responses matched ELIZA's behaviors

We bring to any encounter with other people an immense background of prototypes and mental models of social situations, social types, behavioral scripts, etc. (Lakoff 1987; Rosch 1975; Simmel 1959) These are essential to our ability to socialize, to converse, to make sense of the people around us. Whether we believe that the being with whom we are conversing is intelligent or not is a matter of how well they conform to these expectations.

These mental models allow us to function, to bring our past experience to benefit our present situation. But our way of thinking is also inherently error-prone. We ignore evidence that does not fit into our existing models and we fill in gaps in our experience with beliefs drawn from these models.

For humans, carrying out a conversation is in fact a complex cognitive and affective task. For thousands of years, we have been correct in assuming that anyone with whom we conversed was engaged

in such a process. And indeed, language is the key feature distinguishing us from the other animals. When we speak (or type) with another, we assume that the interplay of words is occurring because our interlocutor is similarly sentient. Even if we are quite aware that the other is a computer, even if we are knowledgeable about its structure and coding, such an exchange seems like it is happening with a sentient being, with personality and selfhood, albeit not quite human.

A computer that can converse may not be mistaken for a human, but it does create a new category of being, the machine conversationalist. Our cultural beliefs about such beings are still evolving, and certainly the technologies themselves are changing rapidly. It is a category that highlights our quick tendency to anthropomorphize, to ascribe human characteristics to a thing that exhibits human-like behavior. And it is an uncomfortable category for many, for it breaks down the barriers between man and machine.

One of the most prized and zealously defended markers of human uniqueness is language. This is certainly true in the popular imagination, where the phrase "dumb animals" refers to all creatures except man, the speaking animal (OED 1989). It is also a strongly held position in the scientific world. Hauser (Hauser 1996)) cites a number of linguists, biologists and other scientists who claim, from their various research perspectives, that human language is unique. "… the possession of articulate speech is the grand distinctive character of man", said T.H. Huxley in 1863. "Language is obviously as different from other animals' communication systems as the elephant's trunk is different from other animals' nostrils", stated Steven Pinker in 1994. Hauser noted that we are far less attached to some other uniquely human features, such as the wearing of clothes or cliff-diving for fun. We see the production of language as an expression of human thought, and thus staunchly defend it.

The value of conversational language as a metric of underlying sentience remains unclear. ELIZA showed that a very simple parser could, if properly framed, seem to be a convincing, if peculiar, human, at least for a brief and highly constrained encounter. 40 years later we are still far from having a machine that can engage in open conversation, via text, with an alert audience, in anything close to a convincing fashion (Loebner 2006)[9] . Yet we are not always an alert audience, vigilantly probing our partners for evidence that they are sentient or human. As Turing noted, "Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks".

But we are entering an era where this convention may no longer hold. Chatterbots – computational descendents of ELIZA – inhabit online chat-rooms, tirelessly striking up conversations with other visitors. Often, they are programmed to present themselves as human, while carrying out a subversive function: disrupting a discussion, marketing a product or extracting personal information. Their existence also casts

[9] This is demonstrated in the annual Loebner Prize competition in which several humans and several computer programs converse (via typing) with a team of judges, who attempt to determine which is human. There is a $100,000 grand prize for the first machine to pass as human in an unrestricted conversation. Thus far, conversations have been restricted to a single topic (known in advance) and no program has yet won, though individual judges have misclassified both machines and humans.

doubt on the human-ness of the truly human participants as angry disputants accuse each other of being robots. A new field of security research is developing to create what are in effect reverse Turing Tests – tests that enable people to prove they are human (von Ahn et al. 2003).

Turing made two predictions in his paper. The first was that in 50 years machines would be able to pass as humans 70 percent of the time in the context of a five minute Imitation Game. This has not yet come to pass, though we can assume that it will happen, though exactly when, and more importantly, with precisely what underlying capabilities, we still do not know for sure. Yet it is likely that it will be a program in mode of ELIZA, designed to cleverly imitate human conversational patterns rather the underlying thought process. Deceiving people is easier than deep imitation.

Turing's second prediction was that "at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expected to be contradicted". Is this true? If so, how did it happen? Why does it matter?

We anthropomorphize machines, using verbs such as "wants' or "likes" to describe actions that have a purely mechanical explanation, but from a behavioral perspective, seem to indicate an intentional behavior. Such machines need not be interactive: our relationship with our cars, for instance, is often highly anthropomorphized. Most people are able to navigate such human/object relationships without truly believing in the sentience of the thing. Very few people today think that such machines "think", in the sense of being conscious beings.

Yet as machines grow more interactive, more (seemingly) intelligent, the definition of machine blurs. Today, the most ambiguous machines are not those pretending to be human, but the ones that claim to be animal. There is, for example, AIBO, the popular robotic pet dog, and Paro, the furry fake seal, designed to nurture and comfort the ill and aged. We perceive these creatures to be in a category different from other non-living objects. Paro's developers cite studies that show that it improves the cognitive functioning of patients with dementia (AIST 2006)[10] – it is touted as having the benefits of a pet without the concerns about cleanliness or the patients' ability to actually care for an animal. Other devices are designed to teach and entertain children, while engaging them both socially and emotionally(Strommen 1998)[11]

[10] The reports of these studies are both enthusiastic and vague.

[11] Some text from an academic paper by the developer of one such product is illustrative of the enthusiastic anthropomorphizing that permeates this field: "ActiMates Barney succeeds as a learning product by conforming to the demands of the situation and to the expectations of his users, just as an intelligent, respectful play partner should. And in this way he suggests an important lesson for future interface designs, especially those based on social conventions. Whether it is an intelligent agent who shops for you, or one that tutors you as you learn, technology interaction is a form of consensual play-acting where both the user and the technology have specific roles to play. Such shared pretend is a form of true partnership and collaboration, and achieving that level of user engagement is the ultimate goal of interface design. It seems particularly fitting that a product designed for our youngest users, for whom

Kahn and Friedman's studies of the owners of robotic dogs conclude that they place them in a new category of being. "This genre comprises artifacts that are *autonomous* (insofar as they initiate action), *adaptive* (act in response to their physical and social environment), *personified* (convey an animal or human persona), and *embodied* (the computation is embedded in the artifacts rather than just in desktop computers or peripherals). If we are correct, then it may be that the English language is not yet well equipped to characterize or talk about this genre." (Kahn et al. 2004)see also (Friedman, Kahn, and Hagman 2003). Turing's prediction about the meaning of words, the implications "thinking" was perhaps not so far off.

This shift in meaning is significant. We are not yet at a stage where we think of machines as human, but we are now at the beginning of an era in which increasingly sophisticated inanimate objects will be part of our everyday life, playing roles that are highly social and emotional (at least from the perspective of the sentient human participant. Even if such machines never achieve even the possibility of consciousness from the perspective of a dispassionate observer armed with the schematics and code that comprise it, the fact that they achieve a believable behavioral sentience makes them a potentially powerful cultural catalyst.

For the technologically optimistic, we will have a world of tireless companions and tutors, nurturers who greet us with unfailing enthusiasm.

Yet it is unclear if we are humanizing machines or mechanizing humans. For Weizenbaum, the acceptance of ELIZA as a therapist was an anti-humanistic statement, evidence that people did not care about the humanity of the other with whom they engaged. Weizenbaum and his family had fled Germany with the rise of the Nazis; he feared that the enthusiastic embrace of machine companionship was symptomatic of the same disregard for the essence of humanity that led to the Holocaust.

In the Ultimatum Game experiment (described in 2.2), people had a neurologically detectable disgust/pain response to receiving low offers and were subsequently willing to pay a cost in order to punish the transgressor. The experimenters ran some rounds of the game with computers rather than humans as the opponents. When the subjects knew the opponent was a machine, the response was quite different: they showed far less activation in the anterior insula than they did when playing against another human. And, their reactions were different: they chose to take the lower amount, instead of paying to punish the opponent (Sanfey et al. 2003)

Our relationships with other people are deeply rooted in the beliefs we have about them. In that experiment, the computer opponent was not endowed with any anthropomorphizing cues. But what if it had been? Certainly if the subjects had been deceived into thinking their opponent was human when in fact it was not we would expect their response to be the same as toward a real human. But what of the more

---

pretend play itself is a way of learning about the world, should remind us of this simple fact." (Strommen 1998).

ambiguous cases, the console that refers to itself as "I", the fur covered monitor? As we live in a world increasingly inhabited by mechanical creatures that nonetheless trigger our social responses, what affect will that have on our expectations and on the ways we learn to treat others?

The behaviorist approach transcends the argument of dualism vs. physicalism, for it allows us address what are becoming pressing issues in social ethics. Does it matter whether a machine could hypothetically think? At some point, yes, if we are concerned about ethical behavior towards machines. But in the meantime, we can assume that whether or machines can potentially think (and whether they will do so in a monistic universe via a different evolution or in a dualistic one via anything from emergent complexity to divine grace) they are not thinking now, and not in the near future - but we are reacting to them as if they are thinking.

Indeed in some ways our current situation, where we know that the machines' sentience is only an act, draws the ethical questions are more sharply. For it forces us to think about when do we care about the inner state of the other, about what they are thinking and feeling, and when do we care only about how they behave. Do we want to be surrounded by people who do care deeply for us – or who just act as if they do?

4. Epilogue: Some history

4.1. History of automata

The fascination with creating automata – man-made objects able to act under their own power and apparent volition – reaches far back in history. What is interesting here is how the mythology, and today the practice, of creating these objects demonstrates prevailing beliefs about both what is needed to make such a thing, whether the touch of the divine or better programming languages, and what is to be feared about doing so.

Western creation myths are tales of divine engineering, telling of a God who creates Man in his own image, though made out of base materials such as metal or clay and given life through Divine power. The tie between creating life and creating objects is explicit in Greek myth of the creation of Pandora, the first woman: it was the god of craft, Hephaestus, who molded her out of clay. A manufactured, engineered Man, created as a copy of a greater being, is the foundation of our culture. This is in contrast to the spiritually organic creation myths from many others cultures, which tell of humans first appearing through birth to the gods (e.g. Japanese) or emerging from the earth (e.g. African Bushmen, Navajo).

Humans, who are imperfect copies of the gods, imitate the divine by attempting to construct their own automatons. Yet it appears that humans may create the form, but only the divine can infuse it with life and consciousness. This occurs in the story of Pygmalion and Galatea. Pygmalion, unimpressed with the local women, created a beautiful statue, Galatea (sleeping love). Galatea was perfect, and Pygmalion gave her gifts and adored her, but she was only a statue until Aphrodite brought her to life. In this myth, reward come to the skilled craftsman for his meticulous and extraordinary work: one of the gods breathed life into his statue and Galatea became his wife and they lived happily ever after.

Similarly, in the Jewish legend of the Golem, the rabbi is able to create the creature's form, but must invoke the divine by inscribing it with the word Emeth (truth) to provide it with life. The early versions of this story are ambivalent about the desirability of creating such a creature, but as the legend entered popular culture the Golem became a figure of evil, a symbol of the evil that ensues from such hubristic pursuits.

The Western creation myths, the story of Pygmalion and the early versions of the Golem legend are spiritually dualistic. There is a material form and there is life and consciousness; only the divine can create the latter. Man's domain is the material world and he is capable of creating wonderful forms, but cannot endow them with life or mind.

The later versions of the Golem myth, including such related tales as Frankenstein's monster, are darker – and less definitively dualistic. Here, man (and perhaps nature, in the form of lightning) is the sole creator of the synthesized being and the tale is about the evil that comes from such endeavors. It is worth noting that Shelley's tale does not say that the monster was inherently evil – but raises the question of whether it became evil because it was not treated as a human would be.

The path to creation has changed radically with the introduction of computers. The creation of the "mind" has become the primary goal, and one that is fundamentally an engineering challenge. And the function of the body has become problematic. For some, it is beside the point, unnecessary for the artificial agent; this was Turing's position and informs a wide range of AI research today (Minsky 2005). Other researchers feel that the body is important, but its purpose is solely to engage the human responder (Breazeal and Scassellati 1998). And finally, others contend that the creation of a mind must be grounded (Harnad 1990) in sensory interaction with the physical world (Roy and Pentland 2002). These endeavors are all rooted in a firm belief in monistic physicalism[12] and in technophilic positivism. In the surrounding culture, however, the descendents of Frankenstein's monsters, transformed into giant computer brains, still wreck havoc (Gibson 1987).

4.2. A note on our place in the history of ideas

In the last five or six hundred years tremendous new scientific understandings of the universe, of nature, and of mankind have uprooted the existing world views (Baumer 1977). Pre-scientific beliefs are generally quite intuitive, based on everyday observation. They evolved to support and conform to the culture's ideals: free from the strict constraints of scientific inquiry, a society with minimal knowledge shapes its world-view based on what it wishes for, its desires and fears. By contrast, worldviews based on scientific observation are constrained by what is: they do not (at least in their ideal form) take into account how we would like things to be, but only what actually exists. The Ptolemaic universe, with its round Earth in the center of an encompassing sphere of heavenly bodies not only matched intuitive observations (we are standing still, the sun is rising and setting around us), it also fit into a world view in which the Earth and its life-forms and especially Man, was the focus and purpose of creation. Our current astronomically correct model, in which the Earth is a rather insignificant rock revolving around a star of no particular distinction which is hurtling through infinite space, is not very intuitive as one looks up to the sky, nor is it nearly as deferentially adulatory regarding man's place in the universe.

Galileo removed Man from the center of the universe. Darwin unseated us from the pedestal of special creation and put humans firmly within a chain of being. Yet the same science that has de-mythologized the position of Man in the universe has simultaneously given humanity enormous

[12] Although Searle has argued that an attempt to create mind in a wholly different type of physical structure must be predicated on an implicit dualism (Searle 1980).

technological powers. We may no longer see ourselves in the center of the universe, but we have been able to build rockets, split atoms and reach the moon. We may be humbled to learn that 96% of the human genome is shared by chimpanzees, yet we now have invented antibiotics, insecticides, polyvinylchloride, and are beginning to invent new creatures from DNA on up.

In the coming century the brain/mind sciences – including neuroscience, cognitive science and artificial intelligence – will be the force behind the next wave of cultural upheaval. For these sciences will change how we think about the mind and our understanding of what thought is. They will change our beliefs about why we behave in particular ways, why we have the ethics we do, and whether we will continue to have them in the future or not. And these sciences will bring about enormous new powers. We will be able not only to measure brain activity, but also to affect thought by stimulating the brain. We will be able to create machines that can behave with great subtlety and sophistication, blurring the boundary between man and machine. And we may be faced with seeing the limitations of our understanding – with having science discover how locked into our limited subjective mind we are[13].

5. References

AIST. 2006. *Paro found to improve brain function in patients with cognition disorders*. Advanced Industrial Science and Technology. Accessed. Available from http://www.aist.go.jp/aist_e/latest_research/2006/20060213/20060213.html.

Allison, Truett. 2001. Neuroscience and morality. *Neuroscientist* 7, no. 5: 360-364.

Baumer, Franklin L. 1977. *Modern european thought*. New York: Macmillan Publishing Co.

Bloom, Paul. 2004. The duel between body and soul. *The New York Times*, Sept 10, 2004.

Breazeal, C and B Scassellati. 1998. Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior*.

Brown, Warren S. 2002. Nonreductive physicalism and soul. *The American Behavioral Scientist* 45, no. 12: 1812-1823.

Carr, Laurie, Marco Iacoboni, Marie-Charlotte Dubeau, John C. Mazziotta, and Gian Luigi Lenzi. 2003. Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences USA*. 100, no. 9: 5497-5502.

Chalmers, David J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2, no. 3: 200-219.

Churchland, Paul M. 1989. Knowing qualia: A reply to jackson. In *A neurocomputational perspective*. Cambridge, MA: MIT Press.

Crick, Francis and Christof Koch. 1998. Consciousness and neuroscience. *Cerebral Cortex* 8: 97-107.

[13] This does not necessarily mean that every person's beliefs will change. There are plenty of people today who are wearing synthetic clothes, flying in airplanes, getting flu shots, whose skepticism about science – from Darwin, geology etc – would, if universal, have precluded the developments that they are enjoying.

_____. 2003. A framework for consciousness. *Nature Neuroscience* 6: 119 - 126.

d'Andrade, Roy. 1987. A folk model of the mind. In *Cultural models in language and thought*, ed. Dorothy Holland and Naomi Quinn:112-148.

Dennett, Daniel C. 1988. Quining qualia. In *Consciousness in modern science*, ed. eds A. Marcel and E. Bisiach: Oxford University Press.

Friedman, Batya, Peter H. Kahn, Jr., and Jennifer Hagman. 2003. Hardware companions? What online aibo discussion forums reveal about the human-robotic relationship. In *Proceedings of the SIGCHI conference on Human factors in computing systems*:273-280. Ft. Lauderdale, Florida, USA: ACM Press.

Gibson, William. 1987. *Count zero*. New York: Ace.

Harnad, Steven. 1990. The symbol grounding problem. *Physica D* 42: 335-346.

Hauser, Marc D. 1996. *The evolution of communication*. Cambridge, MA: MIT Press.

_____. 2006. *Moral minds*. New York: HarperCollins.

Hohwy, Jakob and C Frith. 2004. Can neuroscience explain consciousness? *Journal of consciousness studies* 11, no. 7-8: 180-198.

Hrdy, Sarah Blaffer. 1999. *Mother nature*. New York: Pantheon Books.

Jackson, Frank. 1986. What mary didn't know. *The Journal of Philosophy* 83, no. 5: 291-295.

James, William. 1904. Does 'consciousness' exist? *Journal of Philosophy, Psychology, and Scientific Methods* 1: 477-491.

Kahn, Peter H., Jr., Batya Friedman, Deanne R. Perez-Granados, and Nathan G. Freier. 2004. Robotic pets in the lives of preschool children. In *CHI '04 extended abstracts on Human factors in computing systems*:1449-1452. Vienna, Austria: ACM Press.

Kropotkin, Peter. 1902. *Mutual aid: A factor of evolution*.

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lakoff, George and Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Loebner, Hugh. 2006. *Home page of the loebner prize in artificial intelligence*. Accessed. Available from http://www.loebner.net/Prizef/loebner-prize.html.

Ludlow, Peter, Yujin Nagasawa, and Daniel Stoljar, eds. 2004. *There's something about mary*. Cambridge, MA: MIT Press.

Miller, Geoffrey. F. (2000). New York: Doubleday. 2000. *The mating mind: How sexual choice shaped the evolution of human nature.* New York: Doubleday.

Minsky, Marvin. 2005. Interior grounding, reflection, and self-consciousness. In *Proceedings of an International Conference on Brain, Mind and Society*. Tohoku University, Japan.

Nagel, Thomas. 1974. What is it like to be a bat? *The Philosophical Review, Vol. 83, No. 4. (Oct., 1974), pp. 435-450.* 83, no. 4: 435-450.

Panksepp, Jaak. 2003. Feeling the pain of social loss. *Science* 10, no. 5643: 237 - 239.

Quervain, Dominique J.-F. de, Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. 2004. The neural basis of altruistic punishment. *Science* 305, no. 5688: 1254-8.

Ramachandran, V. S. and W. Hirstein. 1998. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies* 4: 429-57.

Rizzolatt, Giacomo and Laila Craighero. 2004. The mirror-neuron system. *Annual Review of Neuroscience* 27: 169-192.

Rolls, Edmund T. 2004. The functions of the orbitofrontal cortex. *Brain and Cognition* 55, no. 1: 11-29.

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology* 104: 192-233.

Rose, James D. 2002. The neurobehavioral nature of fishes and the question of awareness and pain. *Reviews in Fisheries Science* 10, no. 1: 1-38.

Roy, Deb and Alex Pentland. 2002. Learning words from sights and sounds: A computational model. *Cognitive Science* 26, no. 1: 113-146.

Sanfey, Alan G., James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300, no. 5626: 5626.

Searle, John. R. (1980). 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3, no. 3: 417-457.

Simmel, Georg. 1959. How is society possible? In *Georg simmel, 1858-1918: A collection of essays*, ed. Kurt H. Wolff. Columbus, OH: Ohio State University Press. Original edition, 1908.

Strommen, Erik. 1998. When the interface is a talking dinosaur: Learning across media with actimates barney. In *Proceedings of the SIGCHI conference on Human factors in computing systems*:288-295. Los Angeles, California, United States: ACM Press/Addison-Wesley Publishing Co.

von Ahn, Luis, Manuel Blum, Nicholas J. Hopper, and John Langford. 2003. Captcha: Using hard ai problems for security. In *Proceedings of of Eurocrypt '03*:294-311.