Will robots see?

Stanley A. Klein

Will robots ever *really* see? Imagine a robot 10,000 years in the future. At that time we should finally understand the visual system and be able to build robots whose visual performance surpasses humans'. We will then ask: when a robot claims to be seeing, are its sensations the same as mine? The problem is to connect the first person "feel" of an event (qualia) to the third person performance or explanation of the same event. Neuroscience is gradually making progress in connecting third person behavior (including your consciousness) to the simpler activity and interactions of neurons and molecules. The question to be discussed in this chapter is different. It is how subjective, first person, qualia (my consciousness) fits into the structure of objective science.

This chapter has three parts. First, I examine what it will take to convince me that robots can truly see or feel. I will argue that passing a performance test, such as the Turing test (Turing, 1950), is not sufficient since a performance test does not tell me what the robot is actually *feeling*. I do believe, however, that a robot from 10,000 years in the future should be able to convince me that it really does have the same raw sensations (e.g. the color qualia) as those that I feel. However, even though robotic qualia may someday exist, I claim that there will still be a barrier to reducing the qualia to the robot's circuitry. The preceding sentence with its denial of reductionism would sound quite weird were it not for the existence of another nonreductionistic theory: quantum mechanics. Quantum mechanics, the central theory and framework for all of chemistry (and therefore biology) and almost all of physics, is a dualistic theory with a split between the observer and the observed. The second part of the chapter discusses the nature of the quantum duality. The main thesis of this chapter is that the existence of a consistent duality for specifying the role of the observer provides new grounding for exploring how the mind might be related to the body. Knowing that a self-consistent duality is possible gives one courage to explore the possible connection between the duality of quantum mechanics and the duality of the mind-body problem. The third part of this chapter connects the quantum duality to the qualia duality and the role of the homunculus. The present chapter clarifies and emphasizes points that were made in my more detailed paper (Klein, 1991) on the nature of quantum mechanics and its connection to the mind-body problem.

1. What will it take to convince me that a robot's feeling of redness is similar to mine?

I will be asking whether robots could be given the *human* sensation of redness. This issue is a strong version of the philosophical problem of qualia (qualia are the "raw feelings" of a sensation) and is the essence of the mind-body problem. It is a strong version of the qualia question because I am not merely asking whether robots will feel, but I am asking whether their feelings can be similar to those of humans. Later in this section I will point out that this "strong" version of qualia is in fact easier to establish than the question of can robots have qualia of any sort, not necessarily human.

Asking whether robots will see is equivalent to asking whether they will feel or think. This chapter will at times focus on feeling rather than seeing since, for me, the feel of a throbbing headache is more vivid than seeing red.

Turing (1950) proposed a method to test whether computers can think. His method will be adapted to help us determine whether computers have qualia. In the original Turing test, a human must decide whether his computer terminal is hooked up to a computer or to a human. Let us update the Turing test to the year 12,000 A.D. when futuristic robots look and act like humans (such as the robot "Mr. DATA" in *Star Trek: The New Generation* television series). The amount of knowledge of human experience and human physiology that the robot will require for passing the Turing test is so enormous as to be incomprehensible (Dennett, 1985; French, 1990). However, when we remember that in less than 100 years we have gone from the horse and buggy to massively parallel, *analog*, neural network computers, it is incomprehensible to imagine what technology will be like in 10,000 years. A recent Turing test contest at the Boston Computer Museum showed that for restricted topics of conversation present computers do well (Scientific American, January, 1992), so with an additional 10,000 years, successful robots are quite plausible.

Before deciding whether a robot has feelings it is useful to first ask the same question about humans (other than myself) and animals. There is no way to *prove* that your feeling of redness is the same as mine since a first person experience (qualia) is a private matter (several entertaining stories in Hofstadter and Dennett, 1981 make this quite clear). Most of us do believe, however, that we humans are built pretty much the same, and by empathy we believe that red looks the same to you as it does to me. There is also no doubt in my mind that dogs feel and see. I believe that the pain a dog feels when its paw is stepped on feels similar to my pain when my foot is stepped on. The reason for this belief of mine is that the dog's response to pain is similar to how I would respond. I can easily empathize with the dog's yowl. The question "can robots have feelings" is being asked in the same spirit as the question "do dogs have feelings". I can't prove that a dog's feelings are

similar to mine, but somehow a dog can convince me that they are indeed similar. It might be possible for a robot to do the same.

One might think that a robot whose external appearance is indistinguishable from that of a human might have an easier time convincing me it had feelings than it would be for a dog to convince me. That is not true, at least for me. I would constantly worry that the robot was programmed to be a good actor able to "fake" feelings. Let us suppose that I am engaged in a futuristic Turing test in which my task is to determine whether an individual is a human or a robot; and whether or not it has feelings. Suppose that after ten days of living with it, I would have developed sufficient empathy to believe that its seeing and feeling is similar to mine. Imagine my shock when on the eleventh day I would ask it (or him) how it operates, and it opens up its front panel to reveal wires instead of meat. Suppose that when asked how it was able to respond in such a human-like fashion it tells me that it has an enormous look-up table with facts about how humans would respond. There is little doubt in my mind that even though this robot could pass the Turing test, I would not believe that it had feelings similar to mine, even though from the outside they looked similar. I would think that if the robot did have feelings they would surely be different from my feelings. The situation is similar to the sonar of a bat (Nagel, 1974). Even if I understood the sonar circuitry perfectly I would still have no idea what the sonar feels like to the bat. Thus qualia provide information beyond knowledge of the sonar circuitry. The sonar qualia can not be communicated objectively. That qualia provide something extra is central to the arguments of this paper.

The dog's pain is different from the bat's sonar. Although I don't know the feel of sonar I do believe that a dog's feeling of pain isn't very different from my own pain. The reason is that I have a strong belief that dog's pain circuits are wired up similar to those of humans. Searle (1980) would say that the "meat" of dogs is similar to the "meat" of humans. (In fact much of this argument about whether computers can see is similar to Searle's discussion of whether computers can understand). Since I know that I have feelings and since dogs and I are made of similar meat, I presume that dogs have similar feelings. This is not meant to be a syllogism but rather my attempt to explain why I do indeed think that dogs and I have similar feelings. The sonar of bats, on the other hand, involves a neural system that has no parallel for humans so I have no basis to empathize about that system. Note that the way I have come to these conclusions by using empathy is very different from the type of logical reasoning that I use in the rest of science. This point about different modes of knowledge is compatible with my forthcoming argument for why a dualistic science is needed to handle qualia.

The robot could reconvince me that it had true human feelings if it would sit down with me and give me the following long lecture on human and

robot anatomy and physiology. It would tell me to not be put off by all the wires and silicon. It would say that its lookup tables are not at all like the lookup tables of 20th century computers but rather more like the synaptic lookup tables governing human responses. It would show me its circuit diagrams pointing out the structures that mimic my limbic system, hippocampus, amygdala, frontal lobes, neocortex, etc. It would remind me of how these human neural structures are involved in perceptions and consciousness. Even though the robot's structures were made of silicon, I would gradually become convinced that its brain operated just the same as my brain. It is quite likely that I would again, through empathy, become convinced that the robot had the same feelings and sight as I have.

Let us look at what we have accomplished. I have argued that it should be possible for a robot to someday have the same qualia as mine. Why did I restrict the robot's qualia to be just like mine? The surprising answer is that it is easier to convince me that the robot has my qualia than that it has general qualia. French (1990) has given a provocative analysis of the Turing test and comes to a similar conclusion. He discusses a set of questions whose answers depend intimately on human experience. French concludes:

> "Turing invented the imitation game only as a novel way of looking at the question 'Can machines think?' But it turns out to be so powerful that it is really asking: 'Can machines think exactly like human beings?' As a real test for intelligence, the latter question is significantly less interesting than the former"[1].

The same criticism might be leveled at the present analysis since we are asking whether the robot sees like me rather than whether the robot sees in general. This was done since all that I have to do is become convinced that the robot's circuitry and sensors are isomorphic to mine. I don't have to ask about the general nature of what it means to see. Knowing that I see, and knowing that the robot's visual system works just like mine is enough. I do not believe that this narrower question is "less interesting".

The foregoing argument based on empathy has a critical flaw. Just because I empathize with a dog or robot or other human and believe they feel and see like me, does that make it so? Isn't the situation similar to that of the people I see on the streets near campus who truly believe in the magical power of crystals (remember I live in Berkeley)? Does that mean crystals do indeed have power? (Before answering with a quick NO, one should remember the possible placebo effect whereby crystals woven into one's pants can give one the self-confidence to achieve feats that one would otherwise have avoided. See addendum #2 at the end of this chapter for a comment on the connection between science and religion.) In any case I could be wrong about the robot. The robot might have tricked me, and

[1] French, 1990, p. 64.

my belief that it has feelings does not mean it actually does have feelings. Rather than being a humanoid (a robot with human feelings) the robot might be a zimbo. A zimbo, created by philosophers, is a zombie (a creature with no qualia) that has been upgraded to have access to its internal states. Dennett (1991) argues (not convincingly) that a zimbo's access to its internal states is indistinguishable from consciousness. Let us suppose that the crucial difference between a zimbo and a humanoid is that the zimbo is missing a connection between its frontal lobes and its amygdala. When the robot had the conversation with me about its physiology it neglected to point out its missing critical connections, and in my naivete, I didn't notice. My point is that there is a solid test for whether the robot is a humanoid (establish the equivalence of its anatomy and physiology to a human's). A zimbo may have tricked me into thinking that it is a humanoid, but a more careful investigator would have noticed the missing links and realized that the robot was a zimbo devoid of human qualia.

Notice that in the preceding paragraph I postulated that the critical difference between a humanoid and a zimbo is a particular physiological connection. I did this to be very clear that the presence or absence of qualia is linked to physiology. Past versions of duality had the mental realm decoupled from the physical, mechanistic realm. I want to be very clear that my notions of qualia are solidly grounded in neural activity. Thus, the robot's statement "I have a throbbing headache" could be replaced with no loss by the sentence "My 'throbbing headache neurons' are active". From this identity theory one might conclude that there is nothing special about qualia. A satisfactory theory of qualia must have a place for the "outsider's" third-person point of view that the robot's feeling of redness is identical to its "neural" activity.

In order to provide further clarification on the connection between qualia and physiology it is useful to ask which neurons are hooked up to our visual awareness. A popular guess is that one's visual consciousness must partly reside in primary visual cortex (area V1 of monkey or area 17 of humans and cats). This guess is made because only in V1 does one find the very finest receptive fields and the most detailed retinotopic mapping that seems characteristic of our conscious awareness. Later areas lose the nice retinotopic organization and high resolution of V1. When I say the homunculus partly resides in V1 or that the homunculus is aware of the neural activity in V1, what I really mean is that the human observer can sense the activity of the V1 neurons through introspection (more about the homunculus will come later). A surprising result is associated with this claim. It turns out that one of the clearest structures of V1 anatomy are the ocular dominance columns. The inputs to V1 from the two eyes are not randomly intermixed. One region of cortex gets input mainly from the left eye and an adjacent region gets input mainly from the right eye. The regions alternate approximately every

.5 mm. One would think that if consciousness had access to the activity of V1 neurons one would have awareness of eye of origin. Amazingly enough, most humans do not have conscious access to this information. Thus, if you were looking at a scene and a brief light were flashed to just one eye, you would not be able to tell which eye received the flash. One must be very careful in doing these experiments to be sure that all extraneous cues are removed (such as some light reflected from your nose) since humans are very clever at using subtle cues. Blake and Cormack (1979) claimed that certain humans, called amblyopes, who have lost their binocularity of vision do have access to eye-of-origin information. One might think that by studying the different circuitry of amblyopes one might learn where eye-of-origin awareness resides. However, we showed (Barbeito et al., 1985) that when careful controls were taken with amblyopes they were not able to discriminate eye of origin. We were careful to minimize the sensory cues produced by the distortions that are typically found in the amblyopic eye.

It turns out that the ocular dominance structures are not present in all layers of visual area V1. Although they are easily visible in the input layers of V1, in the output layers there is no distinct segregation of right and left eye selective neurons. Our awareness, therefore, seems to be focussed on these output neurons.This story of our lack of eye-of-origin qualia in spite of clear anatomical structures signalling eye-of-origin information shows how psychophysics and anatomy can be used to pin down the neural substrate of qualia. Few doubt that with continued progress in vision research we will someday have a very pretty story connecting visual qualia with particular neural activity patterns. However, there still remains the problem of the feel of the feeling from an insider's (first person) point of view.

To sharpen the issue of whether qualia are something special it is useful to look at the writings of one of the strongest foes of qualia. Two chapters by Dennett ("Qualia Disqualified", (Dennett, 1991); and "Quining Qualia", (Dennett, 1988)) provide the strongest attacks on the notion of qualia that I have come across. I read the chapters with trepidation, fearing that he would produce devastating arguments showing that qualia made no sense. He provides numerous clever gedanken experiments about how subjective sensations can get mixed up and how they do not contain information beyond what the neurons already know. One of Dennett's most interesting thought experiments is the story of the color scientist, Mary, first proposed by Jackson (1982) to demonstrate the existence of qualia.

> "Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black-and-white room via a black-and-white television monitor. She specializes in neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see... What will happen when Mary is released from her black-and-white room or is given a color television monitor? Will she learn anything or not?

> It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete."[2]

Jackson uses this story to show that there is something extra besides the physical information of associated with knowledge of wavelengths, intensities, and color theory. But Dennett (1991) adds a cute twist to the story:

> "And so, one day, Mary's captors decided it was time for her to see colors. As a trick, they prepared a bright blue banana to present her as her first color experience ever. Mary took one look at it and said "Hey! You tried to trick me! Bananas are yellow, but this one is blue!"[3]

Dennett goes on to say how Mary was able to figure out the true color. He reminds us that she has access to all the physical information that is available. She presumably has some device that can measure the ratios of stimulation of her different cones, enabling her to calculate that her "blue" cones are receiving more than expected stimulation from the banana. By this calculation she figures out that her captors are tricking her. Dennett's purpose in this addendum to the Mary story is to point out that for a clever Mary, the subjective percept does not contain more objective information than what is already in the neural activity. I would agree. Qualia are not necessary for making a decision about whether a banana is blue or yellow. The neural firing rates contain that information. Dennett seems to be missing a crucial point, however. From Mary's point of view there is a big difference between her figuring out that the banana was blue based on her calculation of the relative cone catches and her subjective impression of blue. Simply calculating the relative cone catches would not inform her of the *feel* of blue. Churchland (1986) does, however, remind us of an important caveat. In the original Mary story we are told that she has *unlimited* knowledge of human physiology. Churchland presumes that part of her unlimited knowledge is an ability to imagine or hallucinate any desired brain state. In that case Mary would know the feel of the desired color qualia since she embodies human physiology. However, if Mary didn't have the ability to hallucinate new brain states (or to selectively stimulate particular neurons), then even though she knew all about the physiology of the color system she would not know the feel of red if the only stimuli available were black and white. To my relief, all of Dennett's arguments against qualia are similar to the Mary story: that qualia add no measurable information. My point is that although the qualia information is not measurable, it does add a subjective *feel*.

Philosophers have gone in many circles around this question of how to integrate raw subjective feelings with objective neural activity. Is qualia

[2] Jackson, 1982, p. 128.
[3] Dennett, 1991, p. 399.

nothing other than the neural activity or is it something extra? I suspect that I, too, would have continued to flip-flop on this topic for the rest of my life if I hadn't been aware of that most peculiar scientific theory called quantum mechanics. As will be discussed in the last part of this chapter, quantum mechanics has legitimatized the role of the observer, the homunculus, as being outside the normal reductionistic laws of nature. Quantum mechanics provides a language for having reductionism at the same time as having an observer outside the reductionist framework. It is the natural language for discussing the mind-body problem.

Quantum mechanics is the theory underlying all of chemistry and biology. It is the theory that tells us how solids and liquids and molecules are reduced to atoms, how atoms are reduced to electrons, protons and neutrons and how protons and neutrons are reduced to quarks, gluons and possibly strings. What is surprising about quantum mechanics is that it is a dualistic theory. Quantum mechanics provides an existence proof that an elegant, consistent duality is possible between the observer and the observed. The rest of this chapter is devoted to describing quantum mechanics and pointing out its relevance to the problem of connecting qualia to neurophysiology.

2. The dualistic nature of quantum mechanics

This chapter is not the place to give an introduction to quantum mechanics. I will only point out some aspects of quantum mechanics that are relevant to the mind-body duality. Feynman's book, QED, on the interactions of electrons and light (Feynman, 1985) provides an excellent discussion of the laws of quantum mechanics. Herbert's Quantum Reality (Herbert, 1985) is a well-written book on the interpretations of quantum mechanics written for the layman and Wheeler and Zurek (1983) have collected the most important original articles on the topic of quantum measurement. Further details and references can be found in Klein (1991).

The mind-body duality of Descartes has fallen into disrepute because of the difficulty that philosophers have had in developing a consistent theory of the mind-body split. It is rare to find a respectable book on the mind-body problem that will defend the dualistic nature of mind and body. Rather, philosophers work hard to show that any apparent duality is inconsistent. As an example consider Dennett's treatment of duality in his recent book (Dennett, 1991), in the chapter titled "Why dualism is forlorn". He first claims it violates the law of conservation of energy and is thus in conflict with physics. Then he says:

> "This fundamentally anti-scientific stance of dualism is, to my mind, the most disqualifying feature, and is the reason why in this book I adopt the apparently dogmatic rule that dualism is to be avoided at all costs. It is not that I think I can give a knock-down proof that dualism, in all its forms, is false

or incoherent, but that, given the way dualism wallows in mystery, accepting dualism is giving up."[4]

I include this selection from Dennett because it typifies the attitude of most neurophilosophers. The bashing of dualism is based on the ill-formed dualism of Descartes. Before one attacks dualistic theories in general one had better read up on the dualistic Copenhagen interpretation of quantum mechanics. Dualism is not antiscientific. It is the foundation of the most profound theory in all of science. As will be emphasized in the present chapter the quantum duality is the theory of how the observer and the observed interact. That is exactly what philosophers since Descartes have been seeking. My goal is to clarify the connection between the quantum mechanics duality and the duality relevant to qualia. The main message from quantum mechanics is that it is not necessary to fight duality. Rather one should embrace it. I suspect Descartes would have loved it.

The dualistic nature of quantum mechanics was developed by Niels Bohr and is called the Copenhagen interpretation. Although the Copenhagen interpretation may appear mystical, it is in fact a very pragmatic worldview (Stapp, 1972). The Copenhagen interpretation of quantum mechanics says that the universe must be split into two parts, each of which is governed by very different laws. Above the split is the real world with which we are familiar. Experiments are set up, observations are made, feelings are felt. The observer lives above the split. The world looks almost like the classical world of Newton in which matter exists as particles with definite locations. Below the split the laws are different. No observations are allowed. An amplitude (a complex number) is associated with every path the universe can take. Feynman (1985) provides rules for calculating the amplitude for *each possible path (including paths with particles going backwards in time* in a manner that doesn't violate causality above the split). The total amplitude is obtained by adding up all the individual amplitudes, one for each possible path. The connection between the two halves of the duality is deceptively simple: the probability of any event above the split is given by the square of the total amplitude that was calculated below the split. Several characteristics above and below the split are summarized in Table 1.

The connection that I am making between the mind-body problem and quantum mechanics is not complicated. I am merely pointing out that in quantum mechanics the observer has the very special role of being on the other side of the duality from the underlying laws of nature. I, quite naturally, want to identify the mind, or homunculus, with the observer of quantum mechanics. Other investigators who are exploring the connection between quantum mechanics and the mind have a much more ambitious program. They often identify the mind with the waves beneath the split.

[4] Dennett, 1991, p. 37.

Below split	Above split
to be observed	observer
exact laws (no probabilities)	probabilistic laws
Feynman rules	Classical rules (with some nonlocality)
behaves like waves	behaves like particles
local interactions	some nonlocality
body (neurons causing throbbing headache)	mind (feel of throbbing headache)
determinism	free will

Table 1. *Characteristics above and below the split*

Penrose (1989), for example, believes that consciousness must be understood in terms of a quantum state with quantum transitions. My approach is much more modest. It merely makes legitimate the dualistic language that many previous authors have found convenient to use in describing the mind and brain (Descartes, 1664).

The free will–determinism duality in the above list should not be misunderstood. By free will I do not mean the amount of "freedom" that quantum mechanics allows due to its probabilistic structure. Rather I mean the notion of freedom that comes from having an improved language and status for the homunculus that is provided by physics. An excellent discussion of these issues was presented by Searle (1984). However, since Searle is unhappy with dualisms he acknowledges that

> "when it comes to the question of freedom and determinism, I am — like a lot of other philosophers — unable to reconcile the two."[5]

The quantum duality provides a framework for reconciling these seemingly incompatible ideas.

In order to fully appreciate the Copenhagen Interpretation one must have an understanding of the famous debates between Bohr and Einstein (Bohr, 1949). Their discussions on this topic occurred from 1926, when Heisenberg and Schroedinger developed the quantum rules, until 1935, when Einstein published the paradox (Einstein et al., 1935) that was the precursor to Bell's Theorem. In the Einstein-Bohr debates, Einstein would propose a clever experimental method by which the state of the system could be measured without disturbing the system. Bohr would then think for a few hours or days and then show how Einstein's measuring instruments would disturb the system by exactly the right amount to produce the quantum mechanical uncertainty. Bohr showed that the laws above the split were not deterministic, but rather were probabilistic (see second row of above table). Einstein believed in a "real" universe that existed independent of observers. Bohr believed that an outside observer was needed to make the observed system real since before the observation, the system's properties were not

[5] Searle, 1984, p. 86.

yet decided. With quantum mechanics the act of observation produces a transition converting the wave-like world below the split to the particle- like material world above the split. This transition from the spread-out wave to the localized particle is often called "the collapse of the wave-packet".

Although Bohr "won" the debates, Einstein's cause has not died. There have been many attempts by physicists to develop theories with no need for an outside observer. These theories have the wave-packet collapse at an early well-defined point, well before the human observer. Penrose (1989), for example, believes that quantum gravity (a theory that does not yet exist) will cause the collapse whenever a subsystem has 10^{-5} grams of mass or energy. Most physicists, however, doubt that the duality of quantum mechanics will be changed by gravity. The majority of physicists grudgingly accept the dualistic Copenhagen interpretation.

The modern version of the Bohr-Einstein debate is based on Bell's theorem (Bell, 1965; see Klein, 1991 for details). Bell examined a simple system consisting of two particles with correlated polarizations. We will consider the two particles to be photons (Bell used electrons). He showed that the following three principles can not all be true: reality, locality, and quantum mechanical predictions. By reality one means that each photon had a definite polarization before the measurement was made. Reality means the state of each photon is independent of the observer. Locality means that well-separated particles can not interact simultaneously. For the Bell scenario it means that after the two photons have been separated they can not interact. Bell showed that any theory that is real and local will produce correlated polarizations that disagree with the predictions of quantum mechanics. Bell's theorem with pairs of photons has recently been tested (Aspect et al., 1982) and the quantum predictions were verified. This means that either reality or locality is wrong. My preference is to say that below the split reality is not present and above the split some locality is not present (while causality is maintained).

The substantive issue in the Bohr-Einstein debates and Bell's theorem has an amazing parallel to the debates on whether qualia could be reduced to the activity of underlying neurons. Einstein abhorred a dualistic system. Similarly, most neuroscientists and philosophers consider duality to be ugly, unnecessary and inconsistent. Einstein wanted to show that the phenomenology of what we see in our macroscopic experiments is nothing other than the activity of underlying deterministic particles and fields. Similarly neurophilosophers want to show that our throbbing headaches are nothing other than the activity of our underlying neurons.

An important outcome of the Bohr-Einstein debates was the realization that the new quantum mechanics was the first theory of modern science in which the role of the observer could not be ignored. In a sense, quantum mechanics brings the Copernican revolution full circle and restores the observer to the "center of the universe". In all previous theories the interaction of

the observer could be made arbitrarily small so there was no question about a world "out there" existing independent of observation.

It is thought by some that the collapse of the wave packet is not a real change in the state of the system but rather a change of our *knowledge* of the system. Bell's theorem and Aspect's experiment have eliminated this possibility. Bell showed that it wasn't just knowledge that changed when an observation was made, but rather the state of the system changed. In fancy words, an observation produces an ontological change in the state of the world rather than merely an epistemological change.

In neurophilosophy one also argues about whether the difference between qualia and neural mechanisms is an ontological difference or merely an epistemological difference. Is the feel of my throbbing headache and the neural causes of my throbbing headache merely a difference in how we gather knowledge about the same event? Bell's theorem doesn't often get applied to the mind-body problem, so it is not surprising that neurophilosophers and neuroscientists have been able to argue against a dualistic framework. However, since all agree that the brain is based on quantum mechanics, the observer-observed duality will not be able to be avoided for long.

The big problem and its solution: where is the split? The biggest challenge to any dualistic system is to specify the exact placement of the split and to describe precisely the nature of the interaction between the two halves of the split. Neurophilosophers have not succeeded in developing a clean mind-body duality and for that reason have abandoned the quest. Physicists, von Neumann (1932) in particular, figured out how to do it, in a most elegant manner. On the critical question of where must the split be placed, Von Neumann came up with a most amazing answer: **IT CAN BE PLACED ANYWHERE** (Herbert, 1985). The split is movable. It wasn't easy for nature to make the split movable. The laws of nature below the split (Feynman rules) and the laws above the split (classical mechanics with some nonlocality that doesn't violate causality) must be very special. Also the connection between the two halves (the square of the amplitude below the split equals the probability above the split) is very specific. If any of the laws are changed there is a good chance that the split could no longer move freely. It is not surprising that the vague duality of Descartes didn't succeed. Only the most precise, carefully crafted duality has a chance of being self-consistent.

Physicists usually place the split just below their measuring instruments (bubble chambers and geiger counters). Thus in the physicist's placement, almost everything encountered by humans is above the split. Only microscopic entities are below the split. To the physicist a fancy geiger counter is not needed to make an observation. When a cosmic ray leaves a track in a rock, the rock can be considered to make the observation. When a tree falls in Siberia without a human witness, the tree can be considered to be

real because the tree and the ground can act as the observer. Although a rock can be an observer, it doesn't have feelings since its behavior and inner workings doesn't produce empathy in me. Thus most placements of the split are inappropriate for discussions of the mind-body problem.

3. The quantum duality and the homunculus

For the qualia-neuron (mind-body) duality the split would be placed within my brain. Eugene Wigner is a highly respected physicist who gave a clear discussion of this placement (Wigner, 1961). Since he uses a visual detection task as his gedanken experiment, his article is of interest to psychophysicists. When I talk about placing the split inside my brain, I do not necessarily mean that the split has a spatial location, e.g. placing the limbic neurons on one side and cortical neurons on the other side. The split can be between modes of activity of the brain. Consider, for example, how physicists place the split within a geiger counter that has a needle pointer as the readout device. Only the center of mass mode of the needle is typically placed above the split. If the needle consists of 10^{23} atoms then one mode (just the center of mass mode of motion) would be above the split and $10^{23} - 1$ modes of oscillation would be below the split. For the case of a redness qualia, one would place the redness firing pattern together with the relevant attention and consciousness activity above the split (once they are figured out) and all other neural activity below.

The flexibility of the movable split avoids the paradoxes of previous dualities. If the split is placed within my brain then I am the observer and my qualia are not reducible to neural activity. Other humans would be beneath the split, so from my standpoint, their subjective experiences would be fully reducible to the activity of their neurons. Previous mind-body dualities treated all brains alike so there would have been a problem with reductionism. The quantum mechanical duality, on the other hand, allows me to fully reduce your brain to its neurons, while not reducing my brain.

There is a vast history of physicists writing about the role of human consciousness in quantum mechanics (Wigner, 1961). The main problem with this past research is that the articles by physicists tend to be too technical and are therefore inaccessible to neurophilosophers (Stapp, 1990, is an exception). The evidence that the quantum duality has not had much impact on neurophilosophy is glaring. Consider, for example, the recent books by Patricia Churchland (1986), Paul Churchland (1988) and Dennett (1991) on the mind/brain problem. These books are clear, witty, and intelligent and yet they attack a very old version of duality rather than dealing with the improved quantum duality. The wonderful collection of articles and commentaries by Hofstadter and Dennett (1981) does have some comments on the quantum duality but they do not seem to appreciate how well the quantum observer can serve as the mind-body homunculus. It is because

neurophilosophers do not seem to have caught on to the relevance of the quantum duality that the present article is being written.

Let us now reestablish contact with the discussion at the beginning of this chapter concerning whether robots can have qualia and examine how the quantum duality might fit in. An excellent framework to explore the computational theory of the mind has recently been developed by Searle (1990a). He distinguishes between three positions: Weak AI, Strong AI and Cognitivism.

1 **Weak AI: brain processes (and mental processes) can be simulated on a digital computer**
Searle has no problem accepting the Weak AI position, and I presume neither do most people reading this book on robotic vision. In terms of the duality split, the Weak AI position merely claims that if the split is placed high, then everything below the split, including other people's brains, can be understood in terms of the laws of biology, chemistry and physics. Since few disagree with the Weak AI position, nothing more will be said about it.

2 **Strong AI: the mind is a computer program**
Searle disagrees with this Strong AI position that is held by many neurophilosophers (Searle, 1980; Searle, 1990b) and computer scientists. Based on his definitions of mind (as dealing with meaningful entities) and computer programs (formal manipulations of meaningless symbols), Searle (1984, 1990b) logically deduces that the Strong AI position is false. A similar conclusion comes from considerations of the quantum duality. A computer program is a logical construct so there is no place for the quantum split. The split can only be placed in the real physical world not in an idea or formal program. Thus Strong AI does not allow for an outside observer to provide meaning or to feel qualia. Therefore the mind can not be a computer program. Rather than quibble about whether Searle is being fair in using an abstract computer program rather than a program plus computer it is best to shift the debate from Strong AI to Cognitivism (Searle, 1990a) where the real action is located.

3 **Cognitivism: the brain is a digital computer**
Cognitivism, which claims that the brain operates as a computer (I would include deterministic, analog, neural networks as a form of computer), is the main topic of Searle's recent paper (Searle, 1990a). He argues that the central question "Is the brain a digital computer" is ill-defined because of a tacit assumption requiring a homunculus to be the observer. I believe that the quantum duality provides a framework that removes the problems of Cognitivism related to the homunculus.

Searle's main task in his recent article (Searle, 1990a) is to point out the fallacies of Cognitivism. From my point of view, Searle's main accomplishment in his article is to unintentionally provide a strong argument for the quantum duality approach to the connection between brains and computation. I say unintentionally because Searle would never consider himself a dualist. Searle makes the point that computation is in the eye of the beholder. All the 0's and 1's being manipulated by the computer, the syntax, would be meaningless fluctuations of voltages rather than computations were it not for an outside observer providing the computational interpretation. Searle goes on to show how all of Cognitivism makes a tacit assumption of an outside homunculus. The homunculus can not be reduced to simplified structures within the system without having it disappear with nothing remaining to provide the needed interpretation of symbols.

The Cognitivist need for a homunculus is used by Searle to show that Cognitivism is not meaningful since the tacit need for an outside agent violates the "closed system" basis of Cognitivism. The quantum duality is exactly what is needed to rescue Cognitivism. It allows one to invoke a homunculus that is outside the reductionist laws and yet is compatible with these laws. The beauty of the quantum duality is that the placement of the split is maximally slippery. It can be placed wherever it is needed. As I emphasized before (Klein, 1991) the power of the quantum duality is that it legitimatizes a multiplicity of seemingly incompatible worldviews. Only one worldview at a time is allowed, but they can take turns being true. The uniqueness of the placement has the advantage that it avoids the infinite regress that is often associated with mental models involving the homunculus. By providing a self-consistent framework for the outside observer (the homunculus), quantum duality legitimatizes Cognitivism. The quantum duality allows the subjective feel to have its legitimate place in the description of nature.

Is the homunculus we have been discussing active or passive? The quantum homunculus is often thought of as merely a passive observer. However, in quantum theory the observer does play the most important role of collapsing the wave packet to "determine" which of the many possibilities are to be actualized. Stapp (1990) emphasizes this role as a vital aspect of consciousness. When the quantum split is placed low, near the geiger counters, there is no connection to the mind. However, when it is placed high, near my consciousness machinery, my homunculus becomes the "chooser" of different alternatives. This role of the observer has the flavor of some "freedom" of choice.

Suppose the robot says: "I see red". The meaning of this sentence depends on where the split is placed. If the split is placed between me and the robot (with me as the observer) then the sentence can be totally understood in terms of algorithmic computations by the robot's visual and vocal machinery

connecting the visual stimulation to the sound output. With this placement of the split the robot is not an observer and it would be devoid of qualia, since I am choosing to define "see" as requiring a conscious homunculus as observer. If, however, the split is placed between the robot's consciousness module (or whatever the neural substrate of consciousness will turn out to be) and the rest of its circuitry, then the sentence "I see red" will mean that the *robot* observer is having an experience of red. As discussed in the first part of this paper, if the robot's circuitry was closely matched to my own anatomy and physiology, then I might well come to believe that the robot's percept of red is similar to my own percept. Thus the answer to the question of whether the robot senses the qualia of redness depends on my placement of the split. Some readers will undoubtedly be bothered that a single question can be answered either yes or no. This ambiguity produces a relativistic ontology (Klein, 1991) that may be disturbing at first. For each new placement of the split there are new observers and a new ontology. Quantum physicists are gradually getting accustomed to this strange view of reality. It is time that neurophilosophers also learned about the quantum reality.

It is not easy to get used to the quantum mechanical duality. The different placements of the split can produce bizarre ontologies. I can put the split in Bishop Berkeley's position, just below my own homunculus. This produces the solipsist position in which I am the only observer and the rest of the world exists only if I look. The flexible quantum duality allows the validity of even this most bizarre solipsist position. If the quantum duality can tolerate solipsism it should not be surprising that a lower split placement allows humans (and dogs and robots) to be observers with qualia.

Addendum #1

After most of the above was written, I had a discussion with Searle about some of the issues brought up in this paper. He had one major problem and I had one major surprise. His problem was that he wasn't yet convinced that the duality of quantum mechanics was the same as the duality of the mind-body problem. I believe that he is both right and wrong. For most placements of the quantum split he is correct. When the split is placed low, so that a stone acts as the observer, there is no connection with the mind-body problem. However, I believe the situation is different when the split is placed between the mode of activity of my brain corresponding to what we call "the homunculus" and the rest of my brain and body. In that case I claim that the quantum duality is identical to the mind-body duality. I will continue these discussions with Searle and maybe within a few years he will change his mind.

My big surprise was that Searle felt that when the neural circuitry of consciousness becomes understood (which we both agree is likely in the

next 200 years) the mind-body problem will be resolved. I was surprised because I believe that discovering the neural circuitry is part of third person explanations and is not equivalent to having a framework that has a place for subjective feelings and qualia. Maybe I shouldn't have been surprised since in his excellent chapter on the mind-body problem (Searle, 1984) he is clear that "pains and other mental phenomena just are features of the brain". So it seems that Searle, like most scientists, doesn't believe that there is a mind-body problem. I would, of course, agree that for third person explanations there is no problem in connecting someone's mind to his brain, but I still believe a framework with an observer outside the system is useful for interpreting first person experience. This difference of opinion is directly connected with Searle's rejection of dualism. Our further discussions may help to clear up these misunderstandings. Stay tuned.

Addendum #2

Gerald Edelman has just published (Edelman, 1992) one of the most interesting books on the connection of the mind to biology (I always believe that the most recent book I read is the best one). His viewpoint, like Searle's as just discussed, is that the mind can be understood from biology. Even though Edelman has a brief discussion of quantum physics he does not want to connect the quantum duality to the mind-body duality. Edelman, like most scientists and philosophers, rejects any notion of duality as applied to the mind-body problem. What I think is going on is that there are two types of homunculus that are getting confused. I will call them the biology homunculus and the physics homunculus. Edelman provides an excellent framework for the biology homunculus. It is likely that within 200 years the neural circuitry of consciousness will be understood and the biology homunculus will take its place along with DNA as a "simple" solution to what had been thought to be an "impossible" problem. At that point, most people may well believe that the mind-body problem has been "solved". Indeed, I would agree that the most interesting aspect of the mind-body problem will then be solved. The quest to understand the biology homunculus is in fact, an important reason why I am studying visual perception. However, I say again that gaining an understanding of the biology homunculus will not provide a framework or a language for discussing qualia. What is needed is the physics homunculus, that can act as an observer "outside" of the biological reductionistic brain. The clever quantum duality allows the physics homunculus to be present and observing without disturbing the workings of the biological machinery.

It is worth emphasizing a point made in my earlier paper (Klein, 1991). In one sense the role of the physics homunculus is negligible. It doesn't explain anything about how the brain works. All it does is allow us to not be embarrassed to say that the my raw feel of an experience is outside the

realm of science. Thus, not much meat has been left for the mental half of Descartes' duality. On the other hand, the raw feels of my subjective states are the most important part of my experience. Furthermore, for many people on this planet the physics homunculus has an added attraction. It legitimatizes a spiritual (meaning subjective) realm. Descartes' original duality provided a framework for the coexistence of science and religion. The quantum duality has shown how Descartes' vision can be fleshed out to be a consistent, powerful, and flexible framework for understanding our place in the universe.

Acknowledgements

This work was partially supported by AFOSR grant 89-0238.