

Quantum Theory and the Brain.

Matthew J. Donald

**The Cavendish Laboratory, JJ Thomson Avenue,
Cambridge CB3 0HE, Great Britain.**

e-mail: mjd1014@cam.ac.uk

web site: <http://www.bss.phy.cam.ac.uk/~mjd1014>

May 1988
Revised: May 1989
Appears: Proc. Roy. Soc. Lond. A 427, 43-93 (1990)

Abstract. A human brain operates as a pattern of switching. An abstract definition of a quantum mechanical switch is given which allows for the continual random fluctuations in the warm wet environment of the brain. Among several switch-like entities in the brain, we choose to focus on the sodium channel proteins. After explaining what these are, we analyse the ways in which our definition of a quantum switch can be satisfied by portions of such proteins. We calculate the perturbing effects of normal variations in temperature and electric field on the quantum state of such a portion. These are shown to be acceptable within the fluctuations allowed for by our definition. Information processing and unpredictability in the brain are discussed. The ultimate goal underlying the paper is an analysis of quantum measurement theory based on an abstract definition of the physical manifestations of consciousness. The paper is written for physicists with no prior knowledge of neurophysiology, but enough introductory material has also been included to allow neurophysiologists with no prior knowledge of quantum mechanics to follow the central arguments.

CONTENTS

1. **Introduction.**
 2. **The Problems of Quantum Mechanics and the Relevance of the Brain.**
 3. **Quantum Mechanical Assumptions.**
 4. **Information Processing in the Brain.**
 5. **The Quantum Theory of Switches.**
 6. **Unpredictability in the Brain.**
 7. **Is the Sodium Channel really a Switch?**
 8. **Mathematical Models of Warm Wet Switches.**
 9. **Towards a More Complete Theory.**
- References

1. Introduction.

A functioning human brain is a lump of warm wet matter of inordinate complexity. As matter, a physicist would like to be able to describe it in quantum mechanical terms. However, trying to give such a description, even in a very general way, is by no means straightforward, because the brain is neither thermally isolated, nor in thermal equilibrium. Instead, it is warm and wet — which is to say, in contact with a heat bath — and yet it carries very complex patterns of information. This raises interesting and specific questions for all interpretations of quantum mechanics. We shall give a quantum mechanical description of the brain considered as a family of thermally metastable switches, and shall suggest that the provision of such a description could play an important part in developing a successful interpretation of quantum mechanics.

Our essential assumption is that, when conscious, one is directly aware of definite physical properties of one's brain. We shall try both to identify suitable properties and to give a general abstract mathematical characterization of them. We shall look for properties with simple quantum mechanical descriptions which are directly related to the functioning of the brain. The point is that, if we can identify the sort of physical

substrate on which a consciousness constructs his world, then we shall have a definition of an observer (as something which has that sort of substrate). This could well be a major step towards providing a complete interpretation of quantum mechanics, since the analysis of observers and observation is the central problem in that task. We shall discuss the remaining steps in §9. Leaving aside this highly ambitious goal, however, the paper has three aspects. First, it is a comment, with particular reference to neurophysiology, on the difficulties of giving a fully quantum mechanical treatment of information-carrying warm wet matter. Second, it is a discussion of mathematical models of “switches” in quantum theory. Third, it analyses the question of whether there are examples of such switches in a human brain. Since, ultimately, we would wish to interpret such examples as those essential correlates of computation of which the mind is aware, this third aspect can be seen, from another point of view, as asking whether humans satisfy our prospective definition of “observer”.

The brain will be viewed as a finite-state information processor operating through the switchings of a finite set of two-state elements. Various physical descriptions of the brain which support this view will be provided and analysed in §4 and §6. Unlike most physicists currently involved in brain research (for example, neural network theorists), we shall not be concerned here with modelling at the computational level the mechanisms by which the brain processes information. Instead, we ask how the brain can possibly function as an information processor under a global quantum mechanical dynamics. At this level, even the existence of definite information is problematical.

Our central technical problem will be that of characterizing, in quantum mechanical terms, what it means for an object to be a “two-state element” or “switch”. A solution to this problem will be given in §5, where we shall argue for the naturalness of a specific definition of a switch. Given the environmental perturbations under which the human brain continues to operate normally, we shall show in §7 and §8 that any such switches in the brain must be of roughly nanometre dimension or smaller. This suggests that individual molecules or parts of molecules would be appropriate candidates for such switches. In §6 and §7 we shall analyse, from the point of view of quantum mechanics, the behaviour of a particular class of suitable molecules: the sodium channel proteins. §2 and §3 will be devoted to an exposition of the quantum mechanical framework used in the rest of the paper.

One of the most interesting conclusions to be drawn from this entire paper is that the brain can be viewed as functioning by abstractly definable quantum mechanical switches, but only if the sets of quantum states between which those switches move, are chosen to be as large as possible compatible with the following definition, which is given a mathematical translation in §5:

Definition *A switch is something spatially localized, the quantum state of which moves between a set of open states and a set of closed states, such that every open state differs from every closed state by more than the maximum difference within any pair of open states or any pair of closed states.*

I have written the paper with two types of reader in mind. The first is a neurophysiologist with no knowledge of quantum mechanics who is curious as to why a

quantum theorist should write about the brain. My hope is that I can persuade this type of reader to tell us more about randomness in the brain, about the magnitude of environmental perturbations at neuronal surfaces, and about the detailed behaviour of sodium channel proteins. He or she can find a self-contained summary of the paper in §2, §4, §6, and §7. The other type of reader is the physicist with no knowledge of neurophysiology. This reader should read the entire paper. The physicist should benefit from the fact that, by starting from first principles, I have at least tried to make explicit my understanding of those principles. He or she may well also benefit from the fact that there is no mathematics in the sections which aim to be comprehensible to biologists.

2. The Problems of Quantum Mechanics and the Relevance of the Brain. (This section is designed to be comprehensible to neurophysiologists.)

Quantum theory is the generally accepted physical theory believed to describe possibly all, and certainly most, forms of matter. For over sixty years, its domain of application has been steadily extended. Yet the theory remains somewhat mysterious. At some initial time, one can assign to a given physical object, for example, an electron or a cricket ball, an appropriate quantum mechanical description (referred to as the “quantum state” or, simply, “state” of that object). “Appropriate” in this context means that the description implies that, in as far as is physically possible, the object is both at a fairly definite place and moving at a fairly definite velocity. Such descriptions are referred to by physicists as “quasi-classical states”. The assignment of quasi-classical states at a particular time is one of the best understood and most successful aspects of the theory. The “laws” of quantum mechanics then tell us how these states are supposed to change in time. Often the implied dynamics is in precise agreement with observation. However, there are also circumstances in which the laws of quantum mechanics tell us that a quasi-classical state develops in time into a state which is apparently contrary to observation. For example, an electron, hitting a photographic plate at the end of a cathode ray tube, may, under suitable circumstances, be predicted to be in a state which describes the electron position as spread out uniformly over the plate. Yet, when the plate is developed, the electron is always found to have hit it at one well-localized spot. Physicists say that the electron state has “collapsed” to a new localized state in the course of hitting the plate. There is no widely accepted explanation of this process of “collapse”. One object of this paper is to emphasize that “collapse” occurs with surprising frequency during the operation of the brain.

The signature of “collapse” is unpredictability. According to quantum theory there was no conceivable way of determining where the electron was eventually going to cause a spot to form on the photograph. The most that could be known, even in principle, was the a priori probability for the electron to arrive at any given part of the plate. In such situations, it is the quantum state before “collapse” from which one can calculate these a priori probabilities. That quantum state is believed to provide, before the plate is developed, the most complete possible description of the physical situation. Another goal for this paper is to delineate classes of appropriate quantum

states for the brain at each moment. This requires deciding exactly what information is necessary for a quasi-classical description of a brain.

Now the brain has surely evolved over the ages in order to process information in a predictable manner. The trout cannot afford to hesitate as it rises for the mayfly. Without disputing this fact, however, it is possible to question whether the precise sequence of events in the fish's brain are predictable. Even in those invertebrates in which the wiring diagrams of neurons are conserved across a species, there is no suggestion that a precise and predictable sequence of neural firings will follow a given input. Biologically useful information is modulated by a background of noise. I claim that some of that noise can be interpreted as being of quantum mechanical origin. Although average behaviour is predictable, the details of behaviour can never be predicted. A brain is a highly sensitive device, full of amplifiers and feedback loops. Since such devices are inevitably sensitive to initial noise, quantum mechanical noise in the brain will be important in "determining" the details of behaviour.

Consider once more the electron hitting the photographic plate. The deepest mystery of quantum mechanics lies in the suggestion that, perhaps, even after hitting the plate, the electron is still not really in one definite spot. Perhaps there is merely a quantum state describing the whole plate, as well as the electron, and perhaps that state does not describe the spot as being in one definite place, but only gives probabilities for it being in various positions. Quantum theorists refer in this case to the quantum state of the plate as being a "mixture" of the quantum states in which the position of the spot is definite. The experimental evidence tells us that when we look at the photograph, we only see one definite spot; one element of the mixture. "Collapse" must happen by the time we become aware of the spot, but perhaps, carrying the suggestion to its logical conclusion, it does not happen before that moment.

This astonishing idea has been suggested and commented on by [von Neumann \(1932, §VI.1\)](#), [London and Bauer \(1939, §11\)](#), and [Wigner \(1961\)](#). The relevant parts of these references are translated into English and reprinted in ([Wheeler and Zureck 1983](#)). The idea is a straightforward extension of the idea that the central problem of the interpretation of quantum mechanics is a problem in describing the interface between measuring device and measured object. Any objective physical entity can be described by quantum mechanics. In principle, there is no difficulty with assigning a quantum state to a photographic plate, or to the photographic plate and the electron and the entire camera and the developing machine and so on. These extended states need not be "collapsed". There is only one special case in the class of physical measuring devices. Only at the level of the human brain do we have direct subjective evidence that we can only see the spot in one place on the plate. The only special interface is that between mind and brain.

It is not just this idea which necessitates a quantum mechanical analysis of the normal operation of the brain. It is too widely assumed that the problems of quantum mechanics are only relevant to exceptional situations involving elementary particles. It may well be that it is only in such simple situations that we have sufficiently complete understanding that the problems are unignorable, but, if we accept quantum

mechanics as our fundamental theory, then similar problems arise elsewhere. It is stressed in this paper that they arise for the brain, not only when the output of “quantum mechanical” experiments is contemplated, but continuously.

“Collapse” ultimately occurs for the electron hitting the photographic plate, because the experimenter can only see a spot on a photographic plate as being in one definite place. Even if the quantum state of his retina or of his visual cortex is a mixture of states describing the spot as being visible at different places, the experimenter is only aware of one spot. The central question for this paper is, “What sort of quantum state describes a brain which is processing definite information, and how fast does such a state turn into a mixture?”

One reason for posing this question is that no-one has yet managed to answer the analogous question for spots on a photographic plate. It is not merely the existence of “mixed states” and “collapse” which makes quantum theory problematical, it is the more fundamental problem of finding an algorithmic definition of “collapse”. There is no way of specifying just how blurred a spot can become before it has to “collapse”. There are situations in which it is appropriate to require that electron states are localized to subatomic dimensions, and there are others in which an electron may be blurred throughout an entire electronic circuit. In my opinion, it may be easier to specify what constitutes a state of a brain capable of definite awareness – thus dealing at a stroke with all conceivable measurements - than to try to consider the internal details of individual experiments in a case by case approach.

Notice that the conventional view of the brain, at least among biochemists, is that, at each moment, it consists of well-localized molecules moving on well-defined paths. These molecules may be in perpetual motion, continually bumping each other in an apparently random way, but a snapshot would find them in definite positions. A conventional quantum theorist might be more careful about mentioning snapshots (that after all is a measurement), but he would still tend to believe that “collapse” occurs sufficiently often to make the biochemists’ picture essentially correct. There is still no agreement on the interpretation of quantum mechanics, sixty years after the discovery of the Schrödinger equation, because the conventional quantum theorist still does not know how to analyse this process of collapse. In this paper we shall be unconventional by trying to find the minimum amount of collapse necessary to allow awareness. For this we shall not need every molecule in the brain to be localized.

For most of this paper, we shall be concerned to discover and analyse the best description that a given observer can provide, at a given moment, for a given brain compatible with his prior knowledge, his methods of observation, and the results of his observations. This description will take the form of the assignment of a quantum state to that system. Over time, this state changes in ways additional to the changes implied by the laws of physics. These additional changes are the “collapses”. It will be stressed that the best state assigned by an observer to his own brain will be very different from that which he would assign to a brain (whether living or not) which was being studied in his laboratory.

We are mainly interested in the states which an observer might assign to his own brain. The form of these states will vary, depending on exactly how we assume the

consciousness of the observer to act as an observation of his own brain, or, in other words, depending on what we assume to be the definite information which that brain is processing. We shall be looking for characterizations of that information which provide forms of quantum state for the brain which are, in some senses, “natural”. What is meant by “natural” will be explained as we proceed, but, in particular, it means that these states should be abstractly definable, (that is, definable without direct reference to specific properties of the brain), and it means that they should be minimally constrained, given the information they must carry, as this minimizes the necessity of quantum mechanical collapse.

Interpreting these natural quantum brain states as being mere descriptions for the observer of his observations of his own brain, has the advantage that there is no logical inconsistency in the implication that two different observers might assign different “best” descriptions to the same system. Nevertheless, this does leave open the glaring problem of what the “true” quantum state of a given brain might be. My intention is to leave the detailed analysis of this problem to another work (see §9). I have done this, partly because I believe that the technical ideas in this paper might be useful in the development of a range of interpretations of quantum mechanics, and partly because I wish to minimize the philosophical analysis in this paper. For the present, neurophysiologists may accept the claim that living brains are actually observed in vastly greater detail by their owners than by anyone else, brain surgeons included, so that it is not unreasonable to assume a “true” state for each brain which is close to the best state assigned by its owner. The same assumption may also be acceptable to empirically-minded quantum theorists.

For myself, I incline to a more complicated theory, the truth of which is not relevant to the remainder of the paper. This theory – “the many-worlds theory” – holds, in the form in which it makes sense to me, that the universe exists in some fundamental state ω . At each time t each observer o observes the universe, including his own brain, as being in some quantum state $\sigma_{o,t}$. Observer o exists in the state $\sigma_{o,t}$ which is just as “real” as the state ω . $\sigma_{o,t}$ is determined by the observations that o has made and, therefore, by the state of his brain. Thus, in this paper, we are trying to characterize $\sigma_{o,t}$. The a priori probability of an observer existing in state $\sigma_{o,t}$ is determined by ω . It is because these a priori probabilities are pre-determined that the laws of physics and biology appear to hold in the universe which we observe. According to the many-worlds theory, there is a huge difference between the world that we appear to experience (described by a series of states like $\sigma_{o,t}$) and the “true” state ω of the universe. For example, in this theory, “collapse” is observer dependent and does not affect ω . Analysing the appearance of collapse for an observer is one of the major tasks for the interpreter of quantum theory. Another is that of explaining the compatibility between observers. I claim that this can be demonstrated in the following sense: If Smith and Jones make an observation and Smith observes A rather than B, then Smith will also observe that Jones observes A rather than B. The many-worlds theory is not a solipsistic theory, because all observers have equal status in it, but it does treat each observer separately.

Whatever final interpretation of quantum mechanics we may arrive at, we do assume in this paper, that the information being processed in a brain has definite physical existence, and that that existence must be describable in terms of our deepest physical theory, which is quantum mechanics. Whether the natural quantum brain states defined here are attributes of the observer or good approximations to the true state of his brain, we assume that these natural states are the best available descriptions of the brain for use by the observer in making future predictions. From this assumption, it is but a trifling abuse of language, and one that we shall frequently adopt, to say that these are the states occupied by the brain.

Much of this paper is concerned with discussing how these states change with time. More specifically, it is concerned with discussing the change in time of one of the switch states, a collection of which will form the information-bearing portion of the brain. This discussion is largely at a heuristic (or non-mathematical) level, based on quantum mechanical experience. Of course, in as far as the quantum mechanical framework in this paper is unconventional, it is necessary to consider with particular care how quantum mechanical experience applies to it. For this reason, the pedagogical approach adopted in §6 and §7, is aimed, not only to explain new ideas to biologists, but also to detail suppositions for physicists to challenge.

One central difficulty in developing a complete interpretation of quantum theory based on the ideas in this paper lies in producing a formal theory to justify this heuristic discussion. Such a theory is sketched in §5 and will be developed further elsewhere. The key ingredients here are a formal definition of a switch and a formal definition of the a priori probability of that switch existing through a given sequence of quantum collapses. Some consequences of the switch definition are used in the remainder of the paper, but the specific a priori probability definition is not used. In this sense, the possibility of finding alternative methods of calculating a priori probability, which might perhaps be compatible with more orthodox interpretations of quantum theory, is left open.

3. Quantum Mechanical Assumptions.

(This section is for physicists.)

Four assumptions establish a framework for this paper and introduce formally the concepts with which we shall be working. These assumptions do not of themselves constitute an interpretation of quantum mechanics, and, indeed, they are compatible with more than one conceivable interpretation.

Assumption One *Quantum theory is the correct theory for all forms of matter and applies to macroscopic systems as well as to microscopic ones.*

This will not be discussed here, except for the comment that until we have a theory of measurement or “collapse”, we certainly do not have a complete theory.

Assumption Two *For any given observer, any physical system can best be described, at any time, by some optimal quantum state, which is the state with highest a priori probability, compatible with his observations of the subsystem up to that time.*

(Convention Note that in this paper the word “state” will always mean density matrix rather than wave function, since we shall always be considering subsystems in thermal contact with an environment.)

For the purposes of this paper, it will be sufficient to rely on quantum mechanical experience for an understanding of what is meant by a priori probability. A precise definition is given below in equation 5.6. However, giving an algorithmic definition of this state requires us not only to define “a priori probability”, but also to define exactly what constitutes “observations”. This leads to the analysis of the information processed by a brain. As a consequence, we need to focus our attention, in the first place, on the states of the observer’s brain.

Assumption Three *In the Heisenberg picture, in which operators change in time according to some global Hamiltonian evolution, these best states also change in time. These changes are discontinuous and will be referred to as “collapses”.*

In terms of this assumption and the [previous](#) one, collapse happens only when a subsystem is directly observed or measured. In every collapse, some value is measured or determined. Depending on our interpretation, such a value might represent the eigenvalue of an observable or the status of a switch. Collapse costs a priori probability because we lose the possibilities represented by the alternative values that might have been seen. Thus, the state of highest a priori probability is also the state which is minimally measured or collapsed. This requires a minimal specification of the observations of the observer and this underpins the suggestion in the previous section which led to placing the interface between measuring device and measured object at the mind-brain interface. Nevertheless, a priori probability must be lost continually, because the observer must observe.

Assumption [three](#) is not the same as von Neumann’s “wave packet collapse postulate”. In this paper, no direct link will be made between collapse and the measurement of self-adjoint operators as such. The von Neumann interpretation of quantum mechanics is designed only to deal with isolated and simple systems. I think that it is possible that an interpretation conceptually similar to the von Neumann interpretation, but applying to complex thermal systems, might be developed using the techniques of this paper. I take a von-Neumann-like interpretation, compatible with assumption [one](#), to be one in which one has a state σ_t occupied by the whole universe at time t . Changes in σ_t are not dependent on an individual observer but result from any measurement. Future predictions must be made from σ_t , from the type of collapse or measurement permitted in the theory, and from the universal Hamiltonian. The ideas of this paper become relevant when one uses switches, as defined in [§5](#), in place of projection operators, as the fundamental entities to which definiteness must be assigned at each collapse. The class of all switches, however, is, in my opinion, much too large, and so it is appropriate to restrict attention to switches representing definite information in (human) neural tissue. One would then use a variant of assumption [two](#), by assuming that σ_t is the universal state of highest a priori probability compatible with all observations by every observer up to time t . I do not know how to carry out the details of this programme – which is why I am lead to a many-worlds

theory. However, many physicists seem to find many-worlds theories intuitively unacceptable and, for them, this paper can be read as an attempt to give a definition of “observation” alternative to “self-adjoint operator measurement”. This definition is an improvement partly because it has never been clear precisely which self-adjoint operator corresponded to a given measurement. By contrast, the states of switches in a brain correspond far more directly to the ultimate physical manifestations of an observation.

Assumption Four *There is no physical distinction between the collapse of one pure state to another pure state and the collapse of a mixed state to an element of the mixture.*

This is the most controversial assumption. However, it is really no more than a consequence of assumption [three](#) and of considering non-isolated systems. There is a widely held view that mixed states describe ensembles, just like the ensembles often used in the interpretation of classical statistical mechanics, and that therefore the “collapse” of a mixture to an element is simply a result of ignorance reduction with no physical import. This is a view with which I disagree completely. Firstly, as should by now be plain, the distinction between subjective and objective knowledge lies close to the heart of the problems of quantum mechanics, so that there is nothing simple about ignorance reduction. Secondly, any statistical mechanical system is described by a density matrix, much more because we are looking at only part of the total state of system plus environment, than because the state of the system is really pure but unknown. If we were to try to apply the conventional interpretation of quantum theory consistently to system and environment then we would have to say that when we measure something in such a statistical mechanical system, we not only change the mixed state describing that system, but we also cause the total state, which, for all we know, may well originally have been pure, to collapse.

For an elementary introduction to the power of density matrix ideas in the interpretation of quantum mechanics, see ([Cantrell and Scully 1978](#)). For an example, with more direct relevance to the work of this paper, consider a system that has been placed into thermal contact with a heat bath. Quantum statistical mechanics suggests, that under a global Hamiltonian evolution of the entire heat bath plus system, the system will tend to approach a Gibbs’ state of the form $\exp(-\beta H_s) / \text{tr}(\exp(-\beta H_s))$ where H_s is some appropriate system Hamiltonian. Such a state will then be the best state to assign to the system in the sense of assumption [two](#). Quantum statistical mechanical models demonstrating this scenario are provided by the technique of “master equations”. For a review, see ([Kubo, Toda, and Hashitsume 1985](#), §2.5-§2.7), and, for a rigorously proved example, see ([Davies 1974](#)). These models are constructed using a heat bath which is itself in a thermal equilibrium state, but that tells us nothing about whether the total global state is pure or not. To see this, we can use the following elementary lemma:

lemma 3.1 *Let ρ_1 be any density matrix on a Hilbert space \mathcal{H}_1 , and let H_1 be any Hamiltonian. Let $\rho_1(t) = e^{-itH_1}\rho_1 e^{itH_1}$. Then, for any infinite dimensional Hilbert space \mathcal{H}_2 , there is a pure state $\rho = |\Psi\rangle\langle\Psi|$ on $\mathcal{H}_1 \otimes \mathcal{H}_2$ and a Hamiltonian H such*

that, setting $\rho(t) = e^{-itH}|\Psi\rangle\langle\Psi|e^{itH}$, we have that $\rho_1(t)$ is the reduced density matrix of $\rho(t)$ on \mathcal{H}_1 .

proof Define H by $H(|\psi \otimes \varphi\rangle) = |H_1\psi \otimes \varphi\rangle$ for all $|\psi\rangle \in \mathcal{H}_1$, $|\varphi\rangle \in \mathcal{H}_2$. Let $\rho_1 = \sum_{i \in I} r_i |\psi_i\rangle\langle\psi_i|$ be an orthonormal eigenvector expansion. Choose a set $\{|\chi_i\rangle : i \in I\} \subset \mathcal{H}_2$ of orthonormal vectors and define $\Psi = \sum_{i \in I} \sqrt{r_i} |\psi_i \otimes \chi_i\rangle$. ■

In applying this lemma, I think of ρ_1 as the state of the system plus heat bath, and of \mathcal{H}_2 as describing some other part of the universe. I do not propose this as a plausible description of nature; but it does, I think, suggest that we cannot attach any weight to the distinction between pure and mixed states unless we are prepared to make totally unjustifiable cosmological assumptions. For me, one of the great attractions of the many-worlds interpretation of quantum mechanics is that, because observers are treated separately, it is an interpretation in which collapse can be defined by localized information. Simultaneity problems can thereby be avoided, but the distinction between pure and mixed states is necessarily lost.

One consequence of assumption [four](#) is that the problems to be dealt with in this paper are not made conceptually significantly simpler by the fact that the mathematical descriptions of the brain that we shall employ can almost entirely be expressed in terms of classical, rather than quantum, statistical mechanics. By my view, this means only that, at least at the local level, we are usually dealing with mixtures rather than superpositions, but does not eliminate the problem of “collapse”. Of course, if superpositions never occurred in nature then there might be no interpretation problem for quantum mechanics, but that is hardly relevant. Indeed, it is important to notice that I am not claiming in this paper that the brain has some peculiar form of quantum mechanical behaviour unlike that of any other form of matter. I claim instead that the first step towards an interpretation of quantum mechanics is to analyse the appearance of observed matter, and that a good place to start may be to try to analyse how a brain might appear to its owner. Bohr would have insisted that this means looking for classical (rather than quantum mechanical) behaviour in the brain, but, since I do not believe that Newtonian mechanics has any relevance in neural dynamics, and since I accept assumption [one](#), I have used the word “definite” in place of “classical”.

4. Information Processing in the Brain.

(This section is designed to be comprehensible to neurophysiologists.)

From an unsophisticated point of view, the working of the brain is fairly straightforward. The brain consists of between 10^{10} and 10^{11} neurons (or nerve cells) which can each be in one of two states – either firing or quiescent. The input to the brain is through the firing of sensory neurons in the peripheral nervous system, caused by changes in the external world, and the output is through the firing of motor neurons, which cause the muscles to contract in appropriate response patterns. In-between there is an enormously complex wiring diagram, but, at least as a first approximation, a non-sensory neuron fires only as a result of the firing of other neurons connected to it.

This picture can be refined in every possible aspect, but, leaving aside the details for the moment, we must stress first that the terms in which it is expressed are simply those of a behaviourist view of the brains of others. If we accept, as will be assumed without question in this paper, that we are not simply input-output machines, but that we have some direct knowledge, or awareness, of information being processed in our own brain, then the question arises of what constitutes that information. This question is not answered by merely giving a description of brain functioning. For example, we might consider whether the existence of the physical connections that make up the wiring diagrams forms a necessary part of our awareness, or whether, as will be postulated here, we are only aware of those connections through our awareness of the firing patterns.

In this paper an epiphenomenalist position will be adopted on the mind-body problem. In other words, it will be assumed that mind exists as awareness of brain, but that it has no direct physical effect. The underlying assumption that it is the existence of mind which requires the quantum state of the brain to “collapse”, (because it must be aware of definite information), does not contradict this position, as it will be assumed that the a priori probability of any particular collapse is determined purely by quantum mechanics. Mind only requires that collapse be to a state in which definite information is being processed – it does not control the content of that information. In particular, I assume, that collapse cannot, as has been suggested by [Eccles \(1986\)](#), be directed by the will. Even so, the approach to quantum mechanics taken in this paper does make the epiphenomenalist position much more interesting. Instead of saying that mind must make whatever meaning it can out of a pre-ordained physical substratum, we ask of what sort of substratum can mind make sense.

Finding an interpretation of quantum theory requires us to decide on, or discover, the types of state to which collapse can occur. The aim of this paper is to suggest that, through the analysis of awareness, we can first learn to make that decision by looking at the functioning brain. This will be a matter of supposing that collapse occurs to those states which have just enough structure to describe mentally interpretable substrata. We shall then have a basis in terms of which we can subsequently analyse all collapse or appearance of collapse. Our [assumptions](#) about quantum mechanics imply that it is insufficient to describe a mind merely by the usual hand-waving talk about “emergent properties” arising from extreme complexity, because they imply that the physical information-bearing background out of which a mind emerges is itself defined by the existence of that mind. Thus, we are led to look for simple physical elements out of which that background might be constructed.

One possibility, which we shall refer to as the “neural model”, is that these elements are the firing/quiescent dichotomy of individual neurons. How the information contained in these elements might be made up into that of which we are subjectively aware, remains a matter for hand-waving. What is important instead, is the concrete suggestion that when we try to find quantum states describing a conscious brain, then the firing status of each neuron must be well-defined. In the course of the rest of the paper, we shall provide a whole succession of alternative models of what might be taken to be well-defined in describing a conscious brain. Our first such model,

which we shall refer to henceforth as the “biochemical model” was introduced in §2. It assigns definiteness to every molecular position on, say, an Ångström scale.

The most interesting feature of the neural model is its finiteness. Biologically, even given the caveats to be introduced below, it is uncontroversial to claim that all the significant information processing in a human brain is done through neurons viewed as two-state devices. This implies that all new human information at any instant can be coded using 10^{11} bits per human. Taken together with a rough estimate of total human population, with the quantum mechanical argument that information is only definite when it is observed, and with the idea that an observation is only complete when it reaches a mind, this yields the claim that all definite new information can be coded in something like 10^{21} bits, with each bit switching at a maximum rate of 2000 Hz.

There are two ways of looking at this claim. On the one hand, it says something, which is not greatly surprising, about the maximum rate at which information can be processed by minds. However, on the other hand, it says something quite astonishing about the maximum rate at which new information needs to be added in order to learn the current “collapsed” state of the universe (ignoring extra-terrestrials and animals). Conventional quantum theorists, who would like to localize all molecules (including, for example, those in the atmosphere), certainly should be impressed by the parsimony of the claim. Of course, many important questions have been ignored. Some of these, and, in particular, the question of memory and the details of the analysis of time, are left for another work (see §9). We shall say nothing here about how the information about neural status might be translated into awareness. At the very least this surely involves the addition of some sort of geometrical information, so that, in particular, we can specify the neighbours of each neuron. The information of this kind that we shall choose to add will involve the specification of a space-time path swept out by each neuron. While this will undermine the counting argument just given, I believe that that argument retains considerable validity because it is in the neural switching pattern that most of the brain’s information resides.

The neural model, unfortunately, seems to fail simultaneously in two opposing directions. Firstly, it seems to demand the fixing of far more information than is relevant to conscious awareness. For example, it appears that it is often the rhythm of firing of a neuron that carries biologically useful information, rather than the precise timing of each firing. Indeed, few psychologists would dream of looking at anything more detailed than an overall firing pattern in circuits involving many, many neurons. The lowest “emergent” properties will surely emerge well above the level of the individual neuron.

In my opinion, this first problem is not crucial. We know nothing about how consciousness emerges from its physical substrate. For this paper, it is enough to claim that such a substrate must exist definitely, and to emphasize that it is this definiteness, at any level, which presents a problem for quantum theory. In terms of the amount of superfluous information specified, the neural model is certainly an enormous improvement over the biochemical model.

The more serious problem, however, is that neurons are not, in fact, physically simple. Quantum mechanically, a neuron is a macroscopic object of great complexity. After all, neurons can easily be seen under a light microscope, and they may have axons (nerve fibres) of micron diameter which extend for centimetres. Even the idea of firing as a unitary process is simplistic (see e.g. [Scott 1977](#)). Excitation takes a finite time to travel the length of an axon. More importantly, the excitation from neighbouring neurons may produce only a localized change in potential, or even a firing which does not propagate through the entire cell.

Circumventing this problem, while preserving the most attractive features of the neural model, requires us to find physically simple switching entities in the brain which are closely tied to neural firing. We shall have to make precise, at the quantum theoretical level, the meaning of “physically simple” as well as “switching”. This will occupy the technical sections of this paper, but first, in order to find plausible candidates for our switches, we shall briefly review some neurophysiology, from the usual classical point of view of a biochemist. A useful introductory account of this fascinating subject is given by [Eccles \(1973\)](#).

A resting nerve cell may be thought of as an immersed bag of fluid with a high concentration of potassium ions on the inside and a high concentration of sodium ions on the outside. These concentration gradients mean that the system is far from equilibrium, and, since the bag is somewhat leaky, they have to be maintained by an energy-using pump. There is also a potential difference across the bag wall (cell membrane), which, in the quiescent state, is about -70mV (by convention the sign implies that the inside is negative with respect to the outside). This potential difference holds shut a set of gates in the membrane whose opening allows the free passage of sodium ions.

The first stage in nerve firing is a small and local depolarization of the cell. This opens the nearby sodium gates, and sodium floods in, driven by its electro-chemical gradient. As the sodium comes in, the cell is further depolarized, which causes more distant sodium gates to open, and so a wave of depolarization – the nerve impulse – spreads over the cell. Shortly after opening, the sodium gates close again, and, at the same time, other gates, for potassium ions, open briefly. The resulting outflow of potassium returns the cell wall to its resting potential.

Another relevant process is the mechanism whereby an impulse is propagated from one nerve cell to the next. The signal here is not an electrical, but a chemical one, and it passes across a particular structure - the synaptic cleft – which is a gap of about 25nm at a junction – the synapse – where the two cells are very close, but not in fact in contact. When the nerve impulse on the transmitting cell reaches the synapse, the local depolarization causes little bags (“vesicles”) containing molecules of the transmitter chemical (e.g. acetylcholine) to release their contents from the cell into the synaptic cleft. These molecules then diffuse across the cleft to interact with receptor proteins on the receiver cell. On receiving a molecule of transmitter, the receptor protein opens a gate which causes a local depolarization of the receiver cell. The impulse has been transmitted.

This brief review gives us several candidates for simple two-state systems whose states are closely correlated with the firing or quiescence of a given neuron. There are the various ion gates, the receptor proteins at a synapse, and even the state of the synaptic cleft itself – does it contain neuro-transmitter or not? Here we shall concentrate entirely on the sodium gates.

Note that “neural firing states”, “two-state elements”, and “quantum states” make three different uses of the same word. Keeping “state” for “quantum state”, we shall refer to “neural status” and “switches”.

Sodium gating is part of the function of protein molecules called sodium channels, which have been extensively studied. Their properties as channels allowing the passage of ions, and their role in the production of neural firing are well understood. This understanding constitutes a magnificent achievement in the application of physical principles to an important biological system. I believe that many physicists would enjoy the splendid and comprehensive modern account by Hille (1984). Rather less is known about the detailed molecular processes involved in the gating of the channels, although enough is known to tell us that the channels are considerably more complex than is suggested by simply describing them as being either open or shut. Nevertheless, such a description is adequate for our present purposes, and we shall return to consider the full complexities in §7.

Although the opening and closing of a sodium channel gate is an event that strongly suggests that the neuron of which it forms part has fired, neither event is an inevitable consequence of the other. Nevertheless, it is intuitively clear that the information contained in the open/shut status of the channels would be sufficient to determine the information processing state of the brain, at least if we knew which channel belonged to which neuron. Here I wish to make a deeper claim which is less obviously true.

I shall dignify this claim with a title:

The Sodium Channel Model (first version). *The information processed by a brain can be perfectly modelled by a three dimensional structure consisting of a family of switches, which follow the paths of the brain’s sodium channels, and which open and close whenever those channels open and close.*

We can restate the neural and biochemical models in similar terms:

The Neural Model *The information processed by a brain can be perfectly modelled by a three dimensional structure consisting of a family of switches, which follow the paths of the brain’s neurons, and which open and close whenever those neurons fire.*

The Biochemical Model *The information processed by a brain can be perfectly modelled by a three dimensional structure consisting of ball and stick models of the molecules of the brain which follow appropriate trajectories with appropriate interactions.*

To move from the sodium channel model back to the neural model, one would have to construct neurons as surfaces of coherently opening and closing channels.

Having formulated these models, it is time to analyse the nature of “opening and closing” in quantum theory. Neurophysiologists should rejoin the paper in §6, where the definiteness of the paths of a given channel and the definiteness of the times of its opening and closing will be considered.

5. The Quantum Theory of Switches.

(for physicists)

It is an astonishing fact about the brain that it seems to work by using two-state elements. Biologically, the reason may be that a certain stability is achieved through neurons being metastable switches. By Church’s thesis (see, e.g. Hofstadter 1979), if the brain can be modelled accurately by a computer, then it can be modelled by finite state elements. What is astonishing is that suitable such elements seem so fundamental to the actual physical operation of the brain. It is because of this contingent and empirical fact that it may be possible to use neurophysiology to simplify the theory of measurement. Many people have rejected the apparent complication of introducing an analysis of mind into physics, but it may be that this rejection was unwarranted.

If we are to employ the simplicity of a set of switches, then we have to have a quantum mechanical definition of such a switch. Projection operators, with their eigenvalues of zero and one – the “yes-no questions” of Mackey (1963) – will spring at once to mind. One might be able to build a suitable theory of sodium channels using predetermined projections and defining “measurement”, along the lines suggested by von Neumann, by collapse to the projection eigenvectors. The problem with this option lies with the word “predetermined”. I intend to be rather more ambitious. My aim is to provide a completely abstract definition of sequences of quantum states which would correspond to the opening and closing of a set of switches. Ultimately (see §9 and the brief remarks at the end of this section), having defined the a priori probability of existence of such a sequences of states, I shall be in a position to claim that any such sequence in existence would correspond to a “conscious” set of switches, with an appropriate degree of complexity. For the present, it will be enough to look for an abstract definition of a “switch”. Regardless of my wider ambitions, I believe that this is an important step in carrying out the suggestion of von Neumann, London and Bauer, and Wigner.

Five hypothetical definitions for a switch will be given in this section. Each succeeding hypothesis is both more sophisticated and more speculative than the last. For each hypothesis one can ask:

A) Can sodium channels in the brain be observed, with high a priori probability, as being switches in this sense?

B) Are there no sets of entities, other than things which we would be prepared to believe might be physical manifestations of consciousness, which are sets of switches in this sense, which, with high a priori probability, exist or can be observed to exist, and which follow a switching pattern as complex as that of the set of sodium channels in a human brain?

I claim that any definition of which both A and B were true, could provide a suitable definition for a physical manifestation of consciousness. I also claim that,

given a suitable analysis of a priori probability, both **A** and **B** are true for hypothesis below. Most of this paper is concerned with question **A**. I claim that **A** is not true for hypotheses **I** and **II**, but is true for **III**, **IV**, and **V**. This will be considered in more detail in §6, §7, and §8. I also claim, although without giving a justification in this paper, that **B** is only true for hypothesis **V**.

If one wishes a definition based on predetermined projections, then those projections must be specified. To do this for sodium channels, one would need to define the projections in terms of the detailed molecular structure. This is the opposite of what I mean by an abstract definition. An abstract definition should be, as far as possible, constructed in terms natural to an underlying quantum field theory. This may allow geometrical concepts and patterns of projections, but should avoid such very special concepts as “carbon atom” or “amino acid”.

Hypothesis I *A switch is something spatially localized, which moves between two definite states.*

This preliminary hypothesis requires a quantum theory of localized states. Such a theory – that of “local algebras” – is available from mathematical quantum field theory (Haag and Kastler 1964). We shall not need any sophisticated mathematical details of this theory here: it is sufficient to know that local states can be naturally defined. The two most important features of the theory of local algebras, for our purposes, are, firstly, that it is just what is required for abstract definitions based on an underlying quantum field theory, and secondly, that it allows a natural analysis of temporal change, which is compatible with special relativity. Such local states, it should be emphasized again, will, in general, correspond to density matrices rather than to wave functions. We work always in the Heisenberg picture in which these states do not change in time except as a result of “collapse”.

Technically speaking, local algebra states are normal states on a set of von Neumann algebras, denoted by $\{\mathcal{A}(\Lambda) : \Lambda \subset \mathbb{R}^4\}$, which are naturally associated, through an underlying relativistic quantum field theory (Driessler et al. 1986), with the regions Λ of space-time. $\mathcal{A}(\Lambda)$ is then a set of operators which contain, and is naturally defined by, the set of all observables which can be said to be measurable within the region Λ . For each state ρ on $\mathcal{A}(\Lambda)$ and each observable $A \in \mathcal{A}(\Lambda)$, we write $\rho(A)$ to denote the expected value of the observable A in the state ρ . Thus, formally at least, $\rho(A) = \text{tr}(\rho' A)$ where ρ' is the density matrix corresponding to ρ . ρ is defined as a state on $\mathcal{A}(\Lambda)$ by the numbers $\rho(A)$ for $A \in \mathcal{A}(\Lambda)$. A global state is one defined on the set of all operators. This set will be denoted by $\mathcal{B}(\mathcal{H})$ — the set of all bounded operators on the Hilbert space \mathcal{H} . For example, given a normalized wave function $\psi \in \mathcal{H}$, we define a global state ρ by $\rho(A) = \langle \psi | A | \psi \rangle$. A global state defines states $\rho|_{\mathcal{A}(\Lambda)}$ (read, “ ρ restricted to $\mathcal{A}(\Lambda)$ ”) on each $\mathcal{A}(\Lambda)$ simply by $\rho|_{\mathcal{A}(\Lambda)}(A) = \rho(A)$ for all $A \in \mathcal{A}(\Lambda)$.

Recall the **sodium channel model** from the previous section. We have a family of switches moving along paths in space-time. Suppose that one of these switches occupies, at times when it is open or shut, the successive space-time regions $\Lambda_1, \Lambda_2, \Lambda_3, \dots$. We shall suppose that it is open in Λ_k for k odd, and closed in Λ_k for k even.

We choose these regions so that no additional complete switchings could be inserted into the path, but we do not care if, for example, between Λ_1 and Λ_2 , the switch moves from open to some in-between state and then back to open before finally closing.

In order to represent a switch, the regions Λ_k should be time translates of each other, at least for k of fixed parity. Ignoring the latter refinement, we shall assume that $\Lambda_k = \tau_k(\Lambda)$, $k = 1, 2, \dots$ where Λ is some fixed space- time region and τ_k is a Poincaré transformation consisting of a timelike translation and a Lorentz transformation. We shall also assume that the Λ_k are timelike separated in the obvious order.

While it is of considerable importance that our ultimate theory of “collapse” should be compatible with special relativity, the changes required to deal with general Poincaré transformations are essentially changes of notation, so, for this paper, it will be sufficient to choose the τ_k to be simple time translations. We then have a sequence of times $t_1 < t_2 < \dots$ with $\Lambda_k = \{(x_0 + t_k, \mathbf{x}) : (x_0, \mathbf{x}) \in \Lambda\}$.

Under this assumption, $\mathcal{A}(\Lambda_k)$ is a set of operators related to $\mathcal{A}(\Lambda)$ by $\mathcal{A}(\Lambda_k) = \{e^{it_k H} A e^{-it_k H} : A \in \mathcal{A}(\Lambda)\}$ where H is the Hamiltonian of the total quantum mechanical system (i.e. the universe).

Choose two states ρ_1 and ρ_2 on $\mathcal{A}(\Lambda)$. Suppose that ρ_1 represents an open state and ρ_2 a closed state for our switch. The state σ_k on $\mathcal{A}(\Lambda_k)$ which represents the same state as ρ_1 , but at a later time, is defined by $\sigma_k(e^{it_k H} A e^{-it_k H}) = \rho_1(A)$ for all $A \in \mathcal{A}(\Lambda)$. This yields the following translation of hypothesis I into mathematical language:

Hypothesis II *A switch is defined by a sequence of times $t_1 < t_2 < \dots$, a region Λ of space-time, and two states ρ_1 and ρ_2 on $\mathcal{A}(\Lambda)$. The state of the switch at time t_k is given by $\sigma_k(e^{it_k H} A e^{-it_k H}) = \rho_1(A)$ for $A \in \mathcal{A}(\Lambda)$, when k is odd, and by $\sigma_k(e^{it_k H} A e^{-it_k H}) = \rho_2(A)$ for all $A \in \mathcal{A}(\Lambda)$, when k is even.*

This hypothesis allows the framing of an important question: Is there a single global state σ representing the switch at all times, or is “collapse” required? With the notation introduced above, this translates into: Does there exist a global state σ such that $\sigma|_{\mathcal{A}(\Lambda_k)} = \sigma_k$ for $k = 1, 2, 3, \dots$?

Hypotheses I and II demand that we choose two particular quantum states for the switch to alternate between. This is, perhaps, an inappropriate demand. It is a residue of von Neumann’s idea of definite eigenvectors of a definite projection. As we are seeking an abstract and general definition, using which we shall ultimately claim that our consciousness exists because it is likely that it should, it seems necessary to allow for some of the randomness and imperfection of the real world. In the current jargon, we should ask that our switches be “structurally stable”. This means that every state sufficiently close to a given open state (respectively a given closed state) should also be an open (resp. a closed) state.

The question was raised above of whether it was possible to define a single global state for a switch. In terms of the general aim of minimizing quantum mechanical collapse, a description of a switch which assigned it such a global state would be better than a description involving frequent “collapse”. My preliminary motivation for introducing the requirement of structural stability was that, in order to allow for

variations in the environment of the brain, such stability would certainly be necessary if this goal were to be achieved for sodium channels. Now, in fact, as we shall see in the following sections, there is no way that this goal can be attained for such channels, nor, I suspect, for any alternative physical switch in the brain. Because of this, the concepts with which we are working are considerably less intuitively simple than they appear at first sight. It is therefore necessary to digress briefly to refine our idea of “collapse”.

If we are not prepared to admit structural stability, then we must insist that a sodium channel returns regularly to precisely the same state. However, we cannot simply invoke “collapse” to require this because we are not free to choose the results of “collapse”. In the original von Neumann scenario, for example, we write $\Psi = \sum a_n \psi_n$, expressing the decomposition of a wave function Ψ into eigenvectors of some operator. We may be free to choose the operator to be measured but the a priori probabilities $|a_n|^2$ are then fixed, and each ψ_n will be observed with corresponding probability. If we were required to force a sodium channel to oscillate repeatedly between identical states – pure states in the von Neumann scenario – then, we must choose a set of observation times at each of which we must insist that the channel state correspond either to wave function ψ_1 , representing an open channel, or to wave function ψ_2 , representing a closed channel, or to wave functions ψ_3, \dots, ψ_N , representing intermediate states which will move to ψ_1 or ψ_2 at subsequent observations. We would then lose consciousness of the channel with probability $\sum_{n=N+1}^{\infty} |a_n|^2$. I do not believe that suitable wave functions ψ_1 and ψ_2 exist without the accumulating probability of non-consciousness becoming absurdly high. My grounds for this belief are implicit in later sections of the paper, in which I shall give a detailed analysis of the extent to which normal environmental perturbations act on sodium channels.

Even allowing for variations in our treatment of “collapse”, this sort of argument seems to rule out switching between finite numbers of quantum states in the brain. Instead, we are led to the following:

Hypothesis III *A switch is something spatially localized, the quantum state of which moves between a set of open states and a set of closed states, such that every open state differs from every closed state by more than the maximum difference within any pair of open states or any pair of closed states.*

At the end of this section we shall sketch an analysis of “collapse” compatible with this hypothesis, but first we seek a mathematical translation of it. Denote by U (resp. V) the set of all open (resp. closed) states.

It is reasonable to define similarity and difference of states in terms of projections, both because this stays close to the intuition and accomplishments of von Neumann and his successors, and because all observables can be constructed using projections. Suppose then that we can find two projections P and Q and some number δ such that, for all $u \in U$, $u(P) > \delta$ and, for all $v \in V$, $v(Q) > \delta$. It is natural to insist that P and Q are orthogonal, since our goal is to make U and V distinguishable. We cannot require that we always have $u(P) = 1$ or $v(Q) = 1$, because that would not be stable, but we do have to make a choice of δ . It would be preferable, if possible, to make a universal choice rather than to leave δ as an undefined physical constant.

In order to have a positive distance between U and V , we shall require that, for some $\varepsilon > 0$ and all $u \in U$ and $v \in V$, $u(P) - v(P) > \varepsilon$ and, similarly, $v(Q) - u(Q) > \varepsilon$. Again ε must be chosen.

Finally, we require that U and V both express simple properties. This is the most crucial condition, because it is the most important step in tackling question B raised at the beginning of this section. We shall satisfy the requirement by making the projections P and Q indecomposable in a certain sense. We shall require that, for some η , it be impossible to find a projection $R \in \mathcal{A}(\Lambda)$ and either a pair u_1, u_2 in U with $u_1(R) - u_2(R) \geq \eta$ or a pair v_1, v_2 in V with $v_1(R) - v_2(R) \geq \eta$. If we did not impose this condition, for some $\eta \leq \varepsilon$, then we could have as much variation within U or V as between them. Notice that the choice $\eta = 0$ corresponds to U and V consisting of single points. Thus we require $\eta > 0$ for structural stability.

Finding a quantum mechanical definition for a switch is a matter for speculation. Like all such speculation, the real justification comes if what results provides a good description of physical entities. That said, I make a choice of δ , ε , and η by setting $\delta = \varepsilon = \eta = \frac{1}{2}$. This choice is made natural by a strong and appealing symmetry which is brought out by the following facts:

- 1) $u(P) > u(R)$ for all projections $R \in \mathcal{A}(\Lambda)$ with R orthogonal to P , if and only if $u(P) > \frac{1}{2}$.
- 2) The mere existence of P such that $u(P) - v(P) > \frac{1}{2}$ is sufficient to imply that $u(P) > \frac{1}{2} > v(P)$.

Proving these statements is easy.

Choosing $\varepsilon = \frac{1}{2}$ and $\eta = \frac{1}{2}$ corresponds, as mentioned in the introduction, to making U and V as large as possible compatible with hypothesis III.

Hypothesis IV *A switch in the time interval $[0, T]$ is defined by a finite sequence of times $0 = t_1 < t_2 < \dots < t_K \leq T$, a region Λ of space-time, and two orthogonal projections P and Q in $\mathcal{A}(\Lambda)$. For each $t \in [0, T]$, we denote by σ_t the state of the switch at that time. For later purposes it is convenient to take σ_t to be a global state, although only its restriction to the algebra of a neighbourhood of appropriate time translations of Λ will, in fact, be physically relevant.*

We assume that the switch only switches at the times t_k and that “collapse” can only occur at those times. Thus we require that $\sigma_t = \sigma_{t_k}$ for $t_k \leq t < t_{k+1}$.

For $k = 1, \dots, K$ define σ^k as a state on $\mathcal{A}(\Lambda)$ by

$$\sigma^k(A) = \sigma_{t_k}(e^{it_k H} A e^{-it_k H}) \text{ for } A \in \mathcal{A}(\Lambda). \quad (5.1)$$

(This is not really as complicated as it looks – it merely translates all the states back to time zero in order to compare them.)

The σ^k satisfy

- i) $\sigma^k(P) > \frac{1}{2}$ for k odd,
- ii) $\sigma^k(Q) > \frac{1}{2}$ for k even,
- iii) $|\sigma^k(P) - \sigma^{k'}(P)| > \frac{1}{2}$ and $|\sigma^k(Q) - \sigma^{k'}(Q)| > \frac{1}{2}$ for all pairs k and k' of different parity.
- iv) There is no triple (R, k, k') with $R \in \mathcal{A}(\Lambda)$ a projection and k and k' of equal parity such that $|\sigma^k(R) - \sigma^{k'}(R)| \geq \frac{1}{2}$.

Since the remainder of this section is mathematically somewhat more sophisticated, many physicists may wish to skip from here to §6 on a first reading.

Conditions iii and iv can be translated into an alternative formalism. This is both useful for calculations (see §8), and helps to demonstrate that these conditions are, in some sense, natural. The set of states on a von Neumann algebra \mathcal{A} has a norm defined so that

$$\|u_1 - u_2\| = \sup\{|u_1(A) - u_2(A)| : A \in \mathcal{A}, \|A\| = 1\}. \quad (5.2)$$

It will be shown below (lemma 8.11) that

$$\|u_1 - u_2\| = 2 \sup\{|u_1(P) - u_2(P)| : P \in \mathcal{A}, P \text{ a projection}\}. \quad (5.3)$$

Thus the constraint iii on the distance between U and V is essentially that

$$\text{for all } u \in U, v \in V \text{ we must have } \|u - v\| > 1, \quad (5.4)$$

while the constraint iv on the size of U and V is precisely that for all $u_1, u_2 \in U$ (resp. $v_1, v_2 \in V$) we must have

$$\|u_1 - u_2\| < 1 \text{ (resp. } \|v_1 - v_2\| < 1). \quad (5.5)$$

For completeness, two additional constraints have to be added to our hypothetical definition of a switch. First, we should require that the switch switches exactly K times between 0 and T , so that we cannot simply ignore some of our switch's activity. This requirement is easily expressed in the notation that has been introduced. Second, it is essential to be sure that we are following a single object through space-time. For example, hypothesis IV would be satisfied by a small region close to the surface of the sea, if, through wave motion, that region was filled by water at times t_k for k even and by air at times t_k for k odd. To satisfy this second requirement, we shall demand that the timelike path followed by the switch sweeps out the family of time translates of Λ on which the quantum state changes most slowly. This requires some further notation, and uses the (straightforward) mathematics of differentiation on Banach spaces (for details, see Dieudonné 1969, chapter VIII).

Definition Let (H, \mathbf{P}) be the energy-momentum operator of the universal quantum field theory, and let $y = (y^0, \mathbf{y})$ be a four-vector. Let τ_y denote translation through y . τ_y is defined on space-time regions by $\tau_y(\Lambda) = \{(x^0 + y^0, \mathbf{x} + \mathbf{y}) : (x^0, \mathbf{x}) \in \Lambda\}$ and on quantum states by $\tau_y(\sigma)(A) = \sigma(e^{i(y^0 H - \mathbf{y} \cdot \mathbf{P})} A e^{-i(y^0 H - \mathbf{y} \cdot \mathbf{P})})$. As in (5.1), τ_y maps a state on $\tau_y(\Lambda)$ to one on Λ .

Hypothesis V A switch in the time interval $[0, T]$ is defined by a finite sequence of times $0 = t_1 < t_2 < \dots < t_K \leq T$, a region Λ of space-time, two orthogonal projections P and Q in $\mathcal{A}(\Lambda)$, and a time-like path $t \mapsto y(t)$ from $[0, T]$ into space-time. The state of the switch at time t is denoted by σ_t .

The σ_t satisfy:

1) $\sigma_t = \sigma_{t_k}$ for $t_k \leq t < t_{k+1}$.

2) For $t \in [0, T]$, the function $f(y) = \tau_y(\sigma_t)|\mathcal{A}(\Lambda)$ from space-time to the Banach space of continuous linear functionals on $\mathcal{A}(\Lambda)$ is differentiable at $y = y(t)$, and $\inf\{\|df_{y(t)}(X)\| : X^2 = -1, X_0 > 0\}$ is attained uniquely for

$$X = \frac{dy(t)}{dt}. \text{ (By definition } df_{y(t)}(X) = \lim_{h \rightarrow 0} \frac{f(y(t) + hX) - f(y(t))}{h} \text{.)}$$

- 3) Set $\sigma^k = \tau_{y(t_k)}(\sigma_{t_k})|_{\mathcal{A}(\Lambda)}$. (This is the same as (5.1).) Then
- i) $\sigma^k(P) > \frac{1}{2}$ for k odd,
 - ii) $\sigma^k(Q) > \frac{1}{2}$ for k even,
 - iii) $|\sigma^k(P) - \sigma^{k'}(P)| > \frac{1}{2}$ and $|\sigma^k(Q) - \sigma^{k'}(Q)| > \frac{1}{2}$ for all pairs k and k' of different parity.
 - iv) There is no triple (R, k, k') with $R \in \mathcal{A}(\Lambda)$ a projection and k and k' of equal parity such that $|\sigma^k(R) - \sigma^{k'}(R)| \geq \frac{1}{2}$.
- 4) For each odd (resp. even) $k \in \{1, \dots, K-1\}$, there is no pair $t, t' \in [t_k, t_{k+1}]$ with $t < t'$ (resp. $t' < t$) such that setting $\rho_t = \tau_{y(t)}(\sigma_t)|_{\mathcal{A}(\Lambda)}$ and $\rho_{t'} = \tau_{y(t')}(\sigma_{t'})|_{\mathcal{A}(\Lambda)}$, we have
- i) $\rho_{t'}(P) > \frac{1}{2}$,
 - ii) $\rho_t(Q) > \frac{1}{2}$,
 - iii) $|\rho_{t'}(P) - \sigma^{k'}(P)| > \frac{1}{2}$ and $|\rho_{t'}(Q) - \sigma^{k'}(Q)| > \frac{1}{2}$ for all even k' , and $|\rho_t(P) - \sigma^{k'}(P)| > \frac{1}{2}$ and $|\rho_t(Q) - \sigma^{k'}(Q)| > \frac{1}{2}$ for all odd k' , unless there exists a projection $R \in \mathcal{A}(\Lambda)$ such that either $|\rho_{t'}(R) - \sigma^{k'}(R)| \geq \frac{1}{2}$ for some odd k' , or $|\rho_t(R) - \sigma^{k'}(R)| \geq \frac{1}{2}$ for some even k' .

At the end of the previous section, three possible models of the necessary physical correlates of information processing in the brain were presented. I have no idea how a formalism for calculating a priori probabilities in the biochemical model – the most widely accepted model – might be constructed. In particular, I find insuperable the problems of the so-called quantum Zeno paradox (reviewed by Exner (1985, chapter 2)), and of compatibility with special relativity. For the other models which refer to structures consisting of a finite number of switches moving along paths in space-time, not only is it possible to give an abstract definition of such a structure, using an extension of hypothesis V to N switches, but it is possible to calculate an a priori probability which has, I believe, appropriate properties.

Since the extension to N switches is straightforward, it will be sufficient in this paper to give the a priori probability which I postulate should be assigned to any sequence of states $(\sigma_{t_k})_{k=1}^K$ which satisfies hypothesis V, with switching occurring in regions $\tau_{y(t_k)}(\Lambda)$. These regions will be defined by the brain model that we are using. For example, in the sodium channel model, the space-time regions in which a given channel opens or shuts are defined.

Set $\mathcal{B} = \cup\{\mathcal{A}(\tau_{y(t)}(\Lambda)) : t \in [0, T]\}$. \mathcal{B} is the set of all operators on which the states σ_t are constrained by the hypothesis. Let ω be the state of the universe prior to any “collapse”. Then I define the a priori probability of the switch existing in the sequence of states $(\sigma_{t_k})_{k=1}^K$ to be

$$\text{app}_{\mathcal{B}}((\sigma_{t_k})_{k=1}^K | \omega) = \exp\left\{\sum_{k=1}^K \text{ent}_{\mathcal{B}}(\sigma_{t_k} | \sigma_{t_{k-1}})\right\}, \quad (5.6)$$

where we set $\sigma_{t_0} = \omega$. Here $\text{ent}_{\mathcal{B}}$ is the function defined in (Donald, 1986) and discussed further in (Donald, 1987a).

It is not my intention to discuss in this paper the properties of the function $\text{app}_{\mathcal{B}}$. I merely wish to sketch, in passing, some of the most important possibilities

and difficulties for an interpretation of quantum mechanics based on models of the brain like the [sodium channel model](#). Clearly it is necessary at least to indicate that some means of calculating probabilities can be found. It should be noted that I am not attempting to split the original state (ω) of the universe into a multitude of different ways in which can be experienced. This is how the original von Neumann “collapse” scenario discussed above works. I simply calculate an a priori probability for any of the ways in which ω can be experienced. Using these a priori probabilities one can calculate the relative probabilities of experiencing given results of some planned experiment. I claim that these relative probabilities correspond to those calculated by conventional quantum theory. I hope to publish in due course a justification of this claim and a considerably extended discussion of this entire theory (see §9).

There are also important conceptual questions that could be mentioned. For example, how is one to assign a class of instances of one of these models to a given human being? In particular, in as far as sodium channels carry large amounts of redundant information, can one afford to ignore some of them, and thereby increase a priori probability? I suspect that this particular question may simply be ill-posed, being begged by the use of the phrase “a given human being”, but it emphasizes, once again, that providing a complete interpretation of quantum mechanics is a highly ambitious goal. My belief is that the work of this paper provides interesting ideas for the philosophy of quantum mechanics. I think that it also provides difficult but interesting problems for the philosophy of mind, but that is another story.

6. Unpredictability in the Brain.

(This section is designed to be comprehensible to neurophysiologists.)

It is almost universally accepted by quantum theorists, regardless of how they interpret quantum mechanics, that, at any time, there are limits, for any microsystem, to the class of properties which can be taken to have definite values. That class depends on the way in which the system is currently being observed. For example, returning to the electron striking the photographic plate, it is clear that one could imagine that the electron was in a definite place, near to where the spot would appear, just before it hit the plate. However, it is not possible, consistent with experimentally confirmed properties of quantum theory, to imagine that the electron at that time was also moving with a definite speed. This is very strange. To appreciate the full strangeness, and the extent to which there is experimental evidence for it, one should read the excellent popular account by [Mermin \(1985\)](#).

In this paper, not only is this situation accepted, but the position is even taken that least violence is done to quantum theory by postulating at any time a minimal set of physical properties which are to be assigned definite values. In the [sodium channel model of §4](#), it was proposed to take for these properties the open/shut statuses of the sodium channels of human brains. In this section, we shall consider what this proposal implies about the definiteness of other possible properties of the brain, and, in particular, what the definiteness of sodium channel statuses up to a particular moment implies about the subsequent definiteness of sodium channel statuses.

Sodium channel status, according to the [model of §4](#), involves both a path, which the channel follows, and the times at which the channel opens and shuts. We shall

argue that sodium channel paths cannot be well-localised without frequent “collapse”. We shall also see that we must be more precise about what we mean by a channel opening and shutting, but that here too we need to invoke “collapse”. This section is largely concerned with the general framework of this sort of quantum mechanical description of the brain. Having developed this framework, we shall be ready in the next section for a discussion of the details of possible applications of recent neurophysiological models of the action of sodium channel proteins to the specific quantum mechanical model of a switch, proposed in §5. In this section, we consider the inter-molecular level, leaving the intra-molecular level to the next.

The conceptual difficulty at the heart of this section lies in accepting the idea that what is manifestly definite on, for example, an micrograph of a stained section of neural tissue, need not be definite at all in one’s own living brain. This has nothing to do with the fact that the section is dead, but merely with the fact that it is being looked at. Consider, for example, the little bags of neurotransmitter, the “synaptic vesicles”, mentioned in §4. On an electron micrograph, such vesicles, which have dimensions of order $0.1\ \mu\text{m}$, are clearly localized. However, this does not imply that the vesicles in own’s one brain are similarly localized. The reader who would dismiss this as a purely metaphysical quibble, is, once again, urged to read Mermin’s paper. Perhaps all that makes the electron micrograph definite is one’s awareness of a definite image. The vesicles seen on the micrograph must be localized because one cannot see something which is not localized, and one must see something when one looks. In other words, if you look at a micrograph, then it is not possible for your sodium channels to have definite status unless you are seeing that micrograph as a definite picture. That means that all the marks on the micrograph must, at least in appearance, have “collapsed” to definite positions. This, in turn, makes the vesicles, at least in appearance, “collapse” to definite positions.

In §2, the signature of “collapse” was said to be unpredictability. It turns out, under the [assumptions](#) about quantum theory made in this paper, that the converse is also true, or, in other words, that if something appears to take values which cannot be predicted, then, except at times when it is being observed, it is best described by a quantum state in which it takes no definite value. This means that the appearance of unpredictability in the brain is more interesting than one might otherwise think. One purpose of this section is to review relevant aspects of this topic and to encourage neurophysiologists to tell us more about them.

Unpredictability, of course, is relative to what is known. Absolute unpredictability arises in quantum mechanics because there are absolute limits to what is knowable. What is known and what is predictable depends mainly on how recently and how extensively a system has been observed, or, equivalently, on how recently it has been set up in a particular state. It would not be incompatible with the laws of quantum mechanics to imagine that a brain is set up at some initial time in a quantum state appropriate to what we have referred to above as the [biochemical model](#). In this model, all the atoms in the brain are localized to positions which are well-defined on the Ångström scale. The question then would be how rapidly the atom positions become unpredictable, assuming perfect knowledge of the dynamics. Similarly, the

question for the [sodium channel model](#) is how rapidly the sodium channel paths and statuses become unpredictable after an instant when they alone are known. In reviewing experiments, we must be careful to specify just what is observed and known initially, as well as what is observed finally.

The best quantum mechanical description of a property which is not observed gives that property the same probability distribution as one would find if one did measure its value on every member of a large ensemble of identical systems. The property does not have one real value, which we simply do not know; rather it exists in the probability distribution. Examples clarifying this peculiar idea will be given below. It is an idea which even quantum theorists have difficulty in understanding, and in believing, although most would accept that it is true for suitable microsystems. The [assumptions](#) put forward in this paper are controversial in that they demand that such an idea be taken seriously on a larger scale than is usually contemplated.

Let us first consider what can be said about the sodium channel paths in the membrane (or cell wall) of a given neuron. The “fluid mosaic” model of the cell membrane (see, for example, [Houslay and Stanley 1982](#), §1.5) suggests that many proteins can be thought of as floating like icebergs in the fluid bilayer. Experiments (op. cit. p.83) yield times of the order of an hour for such proteins to disperse over the entire surface of the cell. In carrying out such an experiment, one labels the membrane proteins of one cell with a green fluorescing dye, and those of another with a red fluorescing dye. Then, one forces the two membranes to unite, and waits to see how long it takes for the two dyes to mix. The initial conditions for this experiment, are interesting. One is not following the paths of any individual molecules, but only the average diffusion of proteins from the surface of one cell to the surface of the next. Even so, from this and other experiments, one can predict diffusion coefficients of order $1 - 0.01(\mu\text{m})^2\text{s}^{-1}$ for the more freely floating proteins in a fluid bilayer membrane. These diffusion coefficients will depend on the mass of the protein, on the temperature, and on the composition of the membrane.

In fact, it is not known whether sodium channels do float as freely as the fluid mosaic model would suggest, and there is some evidence to suggest the contrary, at least for some of the channels ([Hille 1984](#), pp 366–369, [Angelides et al. 1988](#)). However, we may be sure that there is some continual and random relative motion. Even if the icebergs are chained together, they will still jostle each other. If we wish to localize our sodium channels on the nanometre scale, as will be suggested in the next section, then the diffusion coefficients given above may well still be relevant, but should be re-expressed as $10^3\text{--}10(\text{nm})^2(\text{ms})^{-1}$. It would seem unlikely to me, that links to the cytoskeleton ([Srinivasan 1988](#)) would be sufficient to hold the channels steady on the nanometre scale. It is however possible that portions of the membrane could essentially crystallise in special circumstances, such as at nodes of Ranvier, or in synaptic structures.

According to the remarks made above, if a sodium channel is not observed, then its quantum state is a state of diffusion over whatever region of membrane it can reach. If, for example, it is held by “chains” which allow it to diffuse over an area of $100(\text{nm})^2$, then it will not exist at some unknown point within that area, but rather

it will be smeared throughout it. Assigning a quantum state to the channel gives a precise mathematical description of the smearing. The laws of quantum mechanics tell us that if we place a channel at a well-defined position in a fluid membrane then, in a time of order the millisecond diffusion time, its quantum state will have become a smeared state. Such a state could only be avoided if we knew the precise positions at all times of all the molecules neighbouring our channel, but these positions too will smear, and on the same timescale (Houslay and Stanley 1982, p 41).

As has been mentioned, it is possible to write down a “quasi-classical” quantum state for an entire brain, corresponding, at one moment, to a description of that brain as it would be given by the [biochemical model](#) with all the atoms in well-localized positions. However, because of all the unpredictable relaxation processes in such a warm wet medium, with relaxation starting at the atomic bond vibration time scale of 10–14s, even such a state will inevitably describe each separate channel as being smeared by the time that its measured diffusion time has elapsed. The message of quantum statistical mechanics is that, in a warm wet environment, floating molecules do not have positions unless they are being observed.

The majority of membrane proteins also appear to spin freely in the plane of the membrane, although they cannot rotate through it. Typical rotational diffusion times are measured at 10 - 1000 μ s (Houslay and Stanley 1982, p 82). In quantum mechanical terms this means that, even if we held the centre of mass fixed, after about 1ms a channel protein which started with a quantum state describing something like a biochemist’s ball and stick model, will be best described by a quantum state which has the molecule rotated in the plane with equal probability through every possible angle.

So far we have only considered the motion of a channel within the cell membrane. This should not be taken to imply that the membrane itself has a well-defined locus. In fact, it is clear that the membrane will spread in position due to collisions with molecules on either side of it. However, this spreading will be much slower than the channel diffusion rate, simply because the membrane is so much larger than the individual channels. More importantly, in adopting the [sodium channel model](#) for a brain, we are asking to know the positions of each channel every time that it opens and shuts. Channels have a surface density of order greater than $100(\mu\text{m})^{-2}$ (Hille 1984, chapter nine), so this is equivalent to observing regularly that density of points of the cell surface. This gives strong constraints on the lateral spreading of individual channels. If we know where all the neighbours of one particular channel are, then the plane within which that one floats will be fairly well determined.

If the neurophysiological reader has not already given up this paper in disgust, then he or she is surely demanding an answer to the question, “How can it be possible that our image of the cell as a collection of well-localized molecules can be superseded, given that, throughout biology, that image has allowed such dramatic progress in understanding?” My answer is that I believe that a more accurate description of a cell, which is not observed, involves describing that cell by a probability distribution of collections of well-localized molecules. Each collection within that distribution can be thought of as developing almost independently – almost precisely as it would

were the cell fixed at a succession of instants to be in the state corresponding to only that collection. After all, the language of biochemistry speaks as if molecules were always well-localized, but it does not reveal the precise path followed by any individual molecule. What matters, and what is measured, is, for example, that on average potassium leaks from a depolarized cell, not that a specific ion follows a specific path through a definite channel. The [biochemical model](#) of a cell is an approximation like a model of the economy in which every consumer is assumed to adopt average behaviour. That such a model can be extremely powerful does not imply that plutocrats do not exist. It is a curious fact that although the image of a cell put forward here is very different from the standard biochemical model, both models are equally compatible with modern cell biology. The purpose of the model that I am proposing is to find a new way of looking at deep, important, and long-standing difficulties in the interpretation of physics. Although my model could be disproved by new information from biology, I do not think that it has any immediate implications for that subject.

There is one essential aspect of this quantum picture of the cell which must be stressed. This is the extent to which the properties of each item are interdependent, so that knowing one property implies constraints on others, or equivalently, observing one property implies constraints on the results of other possible observations. It is a mistake to think of the cell as composed of separately delocalized molecules. For example, if we know where one sodium channel molecule is then we know also where part of the cell membrane is, and we know, without needing a second observation that most of the immediate neighbours of the channel are lipid molecules. This is because the biologically most important properties of the cell are true of every element of the probability distribution which makes up its quantum state. In every such element, every sodium channel has the same amino acid sequence, every sodium channel has both ends inside the cell, and every sodium channel curls up into a similar structure. That those channels are not in identical places in every element of the probability distribution does not vitiate the ability of a neuron to propagate an action potential following every sufficient depolarization.

Turn now to consider the extent to which the times of opening and closing of a channel are well-defined. The most direct experiments involve the use of the astonishing “patch-clamp” technique in which an area of membrane so small that it contains but a single channel is impaled on the end of a micro-pipette. The opening of the channel is then directly detectable as a step increase in current. The results indicate that, following a depolarization of the patch mimicking the propagation of an action potential *in vivo*, the channel does not open once with certainty and then close on the same 0.5 millisecond time scale as the action potential. Instead, the somewhat briefer opening of an individual channel is seen as a random process, and the time course of the sodium current through a comparatively large area of membrane can only be obtained by averaging over a long sequence of records from a single channel.

With this experimental set-up, the status of an individual channel is not predictable from the time course of the potential across the membrane. However, what is measured in these experiments is the current through a single channel which has

been manipulated into circumstances in which it provides information of a kind quite different from that which would have been relevant to the mind to which it might originally have belonged. The observation made by a mind on its own brain may well not be as refined as the observation made by an experimenter on a patch-clamped segment of membrane. It is possible that if sufficient information to reconstruct awareness is contained in firing/non-firing status at the neural level, then consciousness will not observe precisely how a channel conducts current, but rather, whether it is in a quantum state which corresponds to an average state for a channel in a recently depolarized cell, or in a quantum state which corresponds to an average state for a channel in a resting cell.

It is important to realise that there is a much richer class of quantum states for a channel than would be allowed for the corresponding classical biochemical ball and stick model. For example, the ball and stick model must be spatially localized, but the quantum state can be spatially “smeared” as described above. The ball and stick model is simply open or shut, but the quantum state can be, at one instant, both open and shut with equal probability. In the ball and stick model, one must choose the positions of the atoms in the aromatic ring of a phenylalanine residue, but a quantum state can describe them as being in a state of “flipping”.

The ball and stick model was used implicitly in §4 in the first version of the “sodium channel model” for information processing in the brain. It is necessary to refine this in view of this quantum mechanical richness. Quantum states can be assigned to systems by using techniques related to quantum statistical mechanics to find the most likely state given particular prior observations (Donald 1986, 1987a, 1987b). There is a choice of state for a sodium channel in as far as there are choices in the way in which it is assumed to be observed. Given the current flowing through a clamped patch, the quantum state of the channel in the patch must always be either open or shut, like the ball and stick model. However, given only the neural status, or only the potential across the membrane, the most likely quantum state for a channel in an intact neuron will be a state in which the most that can be known, at any time, is a positive (status-, or potential-, dependent) probability for the channel to be open and a complementary and also positive probability for it to be closed. Compared to the channel in the patch, in these latter cases one is given less information, and thus the most likely states have higher a priori probabilities.

Recall that we are seeking the least conditions on the quantum state of the world which will allow brains to process definite information, as this minimizes the necessity of quantum mechanical “collapse”. Allowing that these conditions are that local neural status be definite, the quantum states assigned to sodium channels in vivo will be averaged states compared to those seen with a patch clamp. These average states do follow the potential across a membrane with a time course identical to that of the sodium current across a broad region. In the next section, we shall consider more precise definitions of such “average” states.

A second version of the sodium channel model can now be given:

The Sodium Channel Model (*second version*). *The information processed by a brain can be perfectly modelled by a three dimensional structure consisting of a*

family of switches, which follow the paths of the brain's sodium channels, and which open and shut whenever the corresponding neuron is depolarized locally.

Using this model we still have to answer a question about the extent to which timing in the brain is well-defined, but now it is a question about the timing of local depolarization rather than about the details of individual channel gate movements. In applying this new model, we are considering observations of local neural firing status, imagining that the initial information that we are given is the paths of our sodium channels, the previous pattern of neural firing, and the information about to be delivered to our senses by the external world. We must ask to what extent this initial information is sufficient to predict the subsequent neural firing.

I claim that the precise sequence of timing of switchings in channel status over the whole brain, even when linked to timing in local neural firing, cannot be predicted over any significant duration from any of the possible quantum states which the brain can occupy at any initial moment. I would still make this claim if “precise sequence” were interpreted to mean submillisecond timing, and, even in that case, I would take “significant duration” to mean no more than a few seconds. The justification for this claim is partly that, as will be described below, it is in accordance with the observed behaviour of the brain, and is partly based on experience with quantum states. In general, the rate at which predictable information about particular properties is lost in a quantum state is at least as fast as the rate at which those properties would relax in corresponding situations in classical statistical mechanics. This is a statement which is much too broad (and vague) for mathematical proof, but I think that, at the intuitive level, it should be sufficiently plausible to most quantum theorists, that they would only disbelieve it if they were presented with an explicit construction of a relevant model with demonstrably slower loss of information. What this implies is that “collapse” must occur regularly in every aware brain. As a quantum theorist, I find this statement both so credible, when made, as to need little justification, and surprising, because I have not seen it made before. My deeper purpose in this paper is to delineate a precise model for the nature of collapse in the brain.

Even in a classical biological picture, it is well known to be naive to assume that whether a cell fires or not is completely determined by the impulses incident on it. The entire environment of the cell is relevant. The pattern of local electric fields is related to, but not determined by, the impinging impulse traffic. The density of neurotransmitter in a synaptic cleft depends on the time since the last impulse, and so on. Uncertainty as to where the channels are may lead to uncertainty over whether there are enough of them in a given region of neuron to magnify sufficiently the local depolarization caused by receipt of neurotransmitter from a neighbouring neuron and cause cell firing. The amount of neurotransmitter released into a synaptic cleft as a result of presynaptic depolarization has been measured to be unpredictable (Eccles 1986, Korn and Faber 1987).

As was mentioned in §2, the brain is, in the details of its operation, a very sensitive device. At each stage in neural processing, the firing of subsequent neurons depends on the summation of patterns of firing of earlier neurons. If one neuron gets out of step or misfires, then, while the overall pattern will remain similar, the details,

and, in particular, the precise timings of firings of that neuron's neighbours will be altered. Each subsequent alteration will add to changes already produced, until the final details are totally altered.

These comments are sufficient to demonstrate that the claim that neural firing times are unpredictable on a time scale of seconds, does not conflict with the observations of biologists. Indeed, unpredictability at the level of individual neural firing is not a biological problem. Biologically, the brain is designed to produce results at the level of the general behaviour of the whole animal: running away, eating, curling up and going to sleep, for example. To produce this sort of result, any of an enormous class of patterns of firing would be sufficient. The brain is in many ways designed with redundancy so that errors in operation, or the effects of injury, or even of minor distractions, can be smoothed out.

It remains a matter for speculation to what extent the unpredictability of precise details is ever reflected at the behavioural level. While that trout will certainly take the mayfly, think how many trains of thought lie waiting to be shunted through the brain of an idle human!

One consequence of unpredictability at the individual firing level is that neurophysiologists have found more stable properties, such as short term firing frequencies to be more useful indicators of biologically significant information (Burns 1968). Nevertheless, in order to give a simple mathematical definition of the state of a brain which is abstract, in the sense that the state can be defined without knowing the nature of the information being processed, and without arbitrary choices, such as would be required to define a frequency analyser, it seems necessary to work at a level which, in this sense, is sub-biological. For example, in the [current version](#) of the sodium channel model we require to know precise times of local firings. Merely the huge numbers of firings in the nervous system, each depending on what went before, will make these times rapidly unpredictable to any degree of accuracy. At the crudest level, a cell which through a feedback loop causes itself to fire 100 times per second with an uncertainty in each firing time, due to external noise of 10^{-4} s has a millisecond uncertainty after one second, and a ten millisecond uncertainty after a couple of minutes.

The message of this section is that the brain is not a machine made out of hard little wooden balls connected by strong sticks. Rather it is warm and wet and only has those properties which are imposed on it by observation. The increasingly powerful microscopes that have been brought to bear on the brain have given us a more and more accurate listing of its components. They have also given us an increasingly false picture of how those components exist. A microscope is designed to produce a picture. The "collapse" that this implies is a "collapse" of the elements forming the picture into definite places. Even this apparent "collapse" may be an illusion caused by a genuine "collapse" of the brain of the microscopist into a state of seeing a definite picture. That brain state can be seeing a definite picture without itself containing anything as well localized as the elements of the picture. However, some properties must be definite if a brain is to be aware. Choosing those properties to be definable in terms of the status of sodium channels has turned out to raise difficult questions in

specifying just what is meant by a sodium channel of definite status. Answers to some of these questions may be constrained by new biological information, for example, on the matter of the links holding channels to the cell structure, but this will not alter the essential point of the necessity of frequent “collapse”.

7. Is the Sodium Channel really a Switch?.

(This section is designed to be comprehensible to neurophysiologists.)

The proposal that the interpretation of quantum mechanics can be simplified by analysing brain functioning, depends entirely on the idea that the brain functions through a finite number of switches. In §4, the [neural model](#) was rejected on the grounds that a neuron is so complex. In particular, it seems unlikely that it will ever return to close to the same state twice. It is clear that a sodium channel, which has as its major component a glycopeptide subunit with a molecular weight of roughly 260,000, has, at body temperature, many degrees of freedom. This implies that it too probably never returns to exactly the same state twice. In §5, a definition was given for what it should mean for a quantum system to return to the neighbourhood of a given state, where “the neighbourhood” is a set of states, which is close, in a natural sense, to the original state. Only the long-term judgement of my quantum mechanical peers will tell whether my use of the word “natural” in the last sentence is acceptable. Be that as it may, I have proposed a specific and abstract definition for a quantum mechanical switch. As will be seen in this and the following section, it is possible to estimate the physical dimensions of a system which would satisfy this definition. This will confirm that a whole neuron is indeed far too complex, but a portion of a sodium channel large enough to carry a significant part of a gating apparatus would be suitable. This is a contingent fact, which is already enough to make my definition interesting.

It is in this section that I invoke the most specific properties of the human brain. If some of those properties were other, then my theory would fail. Equally there may be neurophysiological facts of which I am unaware which would make it untenable. However, I would not really expect, at this stage, to find either proof or clear disproof of a theory which might yet be developed in many directions. Thus the work of this section should be seen more as an attempt to place the theory on the continuum between plausibility and implausibility, and to describe in some detail what seems to me at present to be the most plausible implementation and its possible variations.

The vagueness of the description of quantum mechanical “collapse” given in §2 may have left some readers with the impression that there is no problem to be considered because quantum theorists are free to change the state they assign to the world whenever the one that they are working with becomes unsatisfactory. Were this true, then an arbitrarily complex switch could be forced by “collapse” to return to its original state. Of course this is not true, and the reason is that collapse is always, not only a choice of one of the possibilities inherent in the initial state of the world, but also a totally random choice. In other words, to return to our example of an electron hitting a photographic plate, not only must the electron hit the plate, but also we have no choice in where it hits the plate. For each region of the plate there

is a predetermined a priori probability that we could eventually see the spot in that region. If we perform the experiment often enough then, in a corresponding portion of the trials, we shall see the spot in that region.

We have argued that “collapse” occurs frequently in the brain of a normally functioning human. A description of that “collapse” will only be satisfactory if we can describe the uncollapsed quantum state of that brain as allowing a choice of inherent possibilities, each of which describes a “collapsed” brain quantum state. For example, if the state of a sodium channel describes it as being “smeared” by thermal diffusion all over the surface of a neuron, then it is possible, in many ways, to “collapse” the state to one of a set of states, each with the channel localized at a different part of the neuron. Nothing will be left over – there is no probability for the channel to leave the neural surface, nor, leaving aside normal protein turnover, is there a non-negligible probability for it to become something other than a functioning channel. Because of this, the delocalization described in the previous section is not a serious problem. The uncertainty in firing time for a neuron is also not a serious problem. States for each individual channel can be “collapsed” into states with a well-defined phase in the cycle through opening and closing. Each possible phase has a corresponding a priori probability which is equal to the probability of the cell being observed at that point in its cycle.

(As an aside, I should note that I am being somewhat disingenuous here. There is certainly a problem for quantum theorists in giving an algorithmic description of the sort of “collapse” which I am proposing. However, in my opinion, this problem can be solved (see §9 and the remarks at the end of §5), and the heuristics that I am presenting can be justified.)

On the other hand, suppose, as an absurd example, that we decided that a human could only be aware of the opening of a sodium channel if during that opening an even number of sodium ions passed through it. Assuming that there is no physical mechanism which disallows an odd number of ions, there is no way that we can choose to “collapse” the state of a channel into such even parity openings on every occasion on which it opens. Indeed, it is clear that the a priori probability of such an opening is one half. Nevertheless, it would still be possible to suggest the following model of information processing in the brain:

The Sodium Channel Model (*even parity opening version*). *The information processed by a brain can be perfectly modelled by a three-dimensional structure consisting of a family of switches, which follow the paths of the brain’s sodium channels, and which open and close whenever those channels open, allow an even number of sodium ions to pass, and then close.*

Because of the large number of channels in any given neuron, this should still be an adequate model of information processing, even although it only allows us to see a randomly chosen sample of roughly half of the switchings that the first version (§4) of the [sodium channel model](#) allows.

So far, contrary to the warnings given earlier, models of the information processed by a brain have been presented as if there was no question about the existence of

that information. Under a classical theory of physics, one does indeed imagine that the brain exists with some definite structure and has some definite behaviour. One is then only faced with the difficulty of analysing that behaviour. However, when quantum physics is used, one also has to show how that behaviour comes to exist or to be observed to exist. In this paper, I have been seeking to model information processing in the brain by an abstract formalism. I claim that any finite family of switches, moving along definite paths in space-time, and each obeying the definition of a quantum switch given as hypothesis V in §5, could constitute the existence of a mind. The complexity of information processed by that mind will depend on the richness of the pattern of switching performed by the family. In §5, a formal method is given for calculating the a priori probability of existence of any such family. I claim that the only families which have the richness of pattern of switching that a neurophysiologist, using classical physics, would think of the sodium channels of a human brain as having, and which can exist with significant a priori probability, do, in fact, correspond either to the sodium channels of a human brain or to some other similar neural entities. This claim rests on two pillars. One is that there are no other physical entities with switching structures like that of a brain. Arguments for this pillar will be presented elsewhere (see §9). The second pillar is that a brain is a structure which, with high a priori probability, can be observed to exist as a family of switches in a quantum mechanical universe. Justifying this pillar, at the heuristic level, is the central purpose of this paper. To do this, it would be sufficient even to argue for the [even parity version](#) of the sodium channel model, although, of course, the [first](#) and [second](#) versions (in §4 and §6, respectively) have richer structures.

From this point of view, we say that there may be many ways in which a brain can observe itself as a family of switches. Each way corresponds to a model of the kind that we have been presenting. At each switching in one of these models, we can think of “collapsing” the quantum state of the brain into a new state in which the switch concerned has definite status. Such a theory of collapse can only be valid if all the possible futures which might be observed from any stage of the process, exist as possible future collapse states from the state reached at that stage, and have relative probabilities of coming to exist that agree with empirical experience. In other words, we decide, by choosing a model, what constitutes a switch in a brain. The laws of quantum mechanics then assign probabilities to every possible pattern of switching that those switches can carry out over any future duration. Each pattern can be forced into existence by a sequence of “collapses”. Every pattern can be interpreted, we assume, as some particular processing of information in the brain. Our theory is then justified as a step towards an interpretation of quantum mechanics, if there exists some choice of switch for which the resulting patterns of information processing and their probabilities form a satisfactory model for the way in which the world appears to us.

Taking for granted in this paper that the state reached at each collapse is a brain state with sodium channels having well-defined statuses, we shall try, in this section, to analyse the channel quantum states which represent the statuses. We shall be particularly concerned with the question of whether these states can be excessively

altered by the normal range of environmental fluctuations. If they could, then it would be impossible, with the correct relative probabilities, to “collapse” them back into the states they would occupy in the absence of those fluctuations. We shall use as a guide the understanding that the mathematical definition of the a priori probability of a future collapse state should agree with the observable a priori probability of actually experiencing that state in the future.

Up to this point, we have been describing the opening and closing of sodium channels as if it were a simple two-state process. The reality is much more complex. The original model of [Hodgkin and Huxley \(1952\)](#) assigns four mobile parts to each channel. One part is an “inactivation gate”. This closes during the depolarization of the neuron, and helps to bring firing to an end. The other three parts form the “activation gate”, and it is their movement which opens the channel and initiates firing. The Hodgkin-Huxley model provides an excellent description of the time course of neural firing, and has proved to be a useful starting point for the interpretation of the wide range of data that has been collected in the intervening years. As might be expected, these data have shown that the more one probes the sodium channel, the more one learns about new behaviours for new circumstances. There are many reviews detailing these complexities, (e.g. [Armstrong 1981](#); [French and Horn 1983](#); [Aldrich, Corey, and Stevens 1983](#); [Catterall 1986a](#); [Begenisich 1987](#); and [Hille 1984](#), chapter 14). The entire protein has recently been sequenced ([Noda et al. 1984, 1986a, 1986b](#)) and the first models based on the implied atomic structure have now appeared ([Noda et al. 1984, 1986a, 1986b](#); [Kosower 1985](#); [Greenblatt, Blatt, and Montal 1985](#); [Guy and Seetharamulu 1986](#); [Catterall 1986a, 1986b](#); and [Salkoff et al. 1987](#)). A definitive picture of the workings of a sodium channel has yet to emerge, but may not be long in coming now that the atomic structure is known. For this paper, the details of the experimental analyses are not necessary. It is sufficient to work with a general picture, since it will become clear that our definition of a switch can be compatible with a wide range of possibilities. The following general scheme will suffice for our purposes:

A sodium channel is a channel which, when open, allows sodium ions to cross the membrane. It is closed through two distinguishable processes. One of these, corresponding to the activation gate, is responsive to the voltage across the membrane, and opens following depolarization. The other, corresponding to the inactivation gate, closes the channel during depolarization. After the membrane returns to its resting potential, the activation gate closes, the inactivation gate re-opens, and the cycle is complete.

Although this scheme is uncontroversial, it only describes the average response of a channel. From the point of view of the standard biochemical picture of a neuron, some channels may not open at all during a depolarization, because the inactivation gate may have remained shut throughout repolarization. Indeed, a frog preparation has been described in which 50% of the channels have closed inactivation gates at rest ([Hille 1984](#), p.43, quoting [Dodge 1961](#)). Similarly, a channel may open in a resting cell. The opening and closing of these gates constitutes a random process with probabilities varying with the voltage across the membrane. Much attention has been given in the

literature to the question of whether, by assigning enough states to the system, it is possible to describe this process as a Markov process. The consensus seems to be that this can be done, but there is debate on the number of states required. The sixteen states that could be produced from the original Hodgkin-Huxley model by four independently moving particles, can, for example, be reduced by placing constraints on the order in which the particles can move, or, at least at the kinetic level, by indistinguishability between some of the particles.

For the theory of this paper, it is not necessary that there be an underlying Markov process. What is necessary is that we should find some part of the channel and two quantum states for that part, which we shall denote by ρ_1 and ρ_2 . These quantum states must differ significantly. Let U (respectively V) be the set of all states neighbouring ρ_1 (resp. ρ_2). The precise meaning of “differ significantly” and of “neighbouring” is given in §5. Then the following is required to be an adequate model of information processing in the brain (what “adequate” means in this context raises all sorts of philosophical difficulties, but these will not be addressed in this paper):

The Sodium Channel Model (*general version*). *The information processed by a brain can be perfectly modelled by a three dimensional structure consisting of a family of switches, which follow the paths of the brain’s sodium channels, and which open and close whenever the quantum state of the appropriate part of that channel moves from U to V and back.*

It is not necessary that there should be any experimental evidence to show that the quantum state of the channel part invariably moves between U and V in the course of the nerve firing cycle. It is sufficient that the occupation of a state in U be indicative of a resting potential across the cell membrane, and that the occupation of a state in V be indicative of a depolarized membrane, as, in this case, a neuron can be described as at rest if a suitable number of its channel parts pass through states in U , and as firing if a suitable number pass through states in V . Each of the three previous versions of the sodium channel model satisfies this requirement.

Several examples of suitable states for channel parts will now be given. I do not, at present, have sufficient information to choose clearly between these examples, but, given further experimental data, the techniques of the remainder of this paper might allow such a choice to be made eventually. It should be remembered also that sodium channels are not the only possible switches in the brain. This multiplicity of possibilities is not something which I see as a failure in the theory, although it does perhaps raise further philosophical questions. I see it as a motivation both for defining the physical manifestation of mind in abstract terms and for a many-worlds approach to quantum theory. I intend ultimately to interpret all ways in which a human brain can be analysed as a family of quantum switches as being ways in which a mind observes that brain, but the purpose of this paper is to show, for its own independent interest, that there is at least one way.

Example 1: Average States. ρ_1 (resp. ρ_2) is the thermal equilibrium quantum state of a small volume in a piece of membrane, containing an immobilized channel, which is held at the resting (resp. firing) potential and at body temperature.

“Immobilized” here can be chosen to mean that the surrounding lipid is taken to be in a crystalline phase. Alternatively, average states could be constructed by releasing a channel from a fixed position and averaging over the state reached after a suitably short diffusion time. The precise definition is irrelevant, because we shall only really be interested in the sets U and V of near-by states. This example corresponds to the second version (§6) of the sodium channel model.

Example 2: Inactivation Gate States. ρ_1 (resp. ρ_2) is the average quantum state of a small volume containing an open (resp. closed) inactivation gate, or some part of it.

This example depends on the assumption that the inactivation gate is a simpler and more switch-like structure than the rest of the channel. It is not clear that this is so, although it is suggested by the Hodgkin-Huxley theory. A molecular model for the inactivation gate is proposed in (Salkoff et al. 1987). If human neurons are like the frog preparation described above in which 50% of the inactivation gates are closed at rest, then the interpretation of the corresponding information processing model would be like that of the *even parity* version of the sodium channel model, in that each gate will respond to only roughly half of all neural firings. However, while the atomic structure of a frog sodium channel is probably very similar to that of a human, the temperatures at which they operate are very different, and so the dynamics are also likely to differ.

Example 3: Sliding Helix States. *The most explicit molecular models of the sodium channel suggest that the activation gate comprises four membrane spanning helices, which, on opening, undergo screw-like motions, involving twists of about 60° and displacements of about 0.5nm away from the cell (Catterall 1986a, 1986b, Guy and Seetharamulu 1986). ρ_1 (resp. ρ_2) can be assigned the average quantum state of a small volume containing part of one of these undisplaced (resp. displaced) helices and part of the fixed background against which it moves.*

Example 4: Tight Shut and Wide Open States. ρ_1 (resp. ρ_2) is the average quantum state of a small volume in a piece of membrane containing an immobilized channel which would be observed to be in a fully shut (resp. fully open) configuration.

This example depends on the idea, which is widespread in the literature, that a channel may have several open states and certainly has several closed states, but that only one of these states is fully open, and only one is maximally closed. In terms of examples 2 and 3, the fully open state will have all helices displaced and the inactivation gate open, while the maximally closed state will have all helices undisplaced and inactivation gate also open. As in example 2, it is possible, due to the sluggishness of the inactivation gate, that, on average, the maximally closed state is not visited between every neural firing, especially during a period of rapid firing.

There are two complementary aspects that must be considered in testing whether these examples can satisfy the definition given in §5 for a quantum switch. One question is whether the states ρ_1 and ρ_2 are sufficiently different, and the other is whether the quantum state of a channel can, with high probability, be “collapsed”

regularly into neighbourhoods of these states. In the remainder of this section, we shall discuss the observational evidence bearing on these questions, leaving the detailed translation into mathematics for the next section.

The definition of “sufficiently different” that has been adopted can be translated as saying, roughly, that ρ_1 and ρ_2 are sufficiently different if there could exist a means of observing each state, within the volume on which that state is defined, which would allow them to be distinguished unambiguously on more than half of all observations. This is a very weak definition of difference. The states in examples [two](#), [three](#), and [four](#) are almost obviously sufficiently different, as long as the volume in question is taken to be large enough to contain some of the molecular bonds that will necessarily change shape or relative position as a result of the motion being considered. It is possible, however, that the states in the [first](#) example may not differ by the amount required. This possibility would arise, for example, if, regardless of potential, as many as half the channels in a nerve were inactivated at any time, because then when one observed the average state of a channel, one would usually come up with an inactivated channel, whatever the potential across the membrane. In this case, one would be forced to abandon example [one](#), or perhaps to modify it as follows:

Example 5: Average States at Fixed Cycle Time. ρ_1 (resp. ρ_2) is the thermal equilibrium quantum state of a small volume in a piece of membrane, containing an immobilized channel, which has been held at the resting potential for some specified time (resp. has been depolarized for some specified time).

The specified times in this example are to be chosen to allow the average channel in ρ_1 (resp. ρ_2) time to close (resp. open).

Another example in which ρ_1 and ρ_2 seem not to be sufficiently different is the following:

Example 6: Average Membrane States. ρ_1 (resp. ρ_2) is the thermal equilibrium quantum state of a small volume in a piece of membrane, which is held at the resting (resp. firing) potential and at body temperature.

Channels are sparsely distributed over the membrane, so there is little probability of hitting a channel in an average piece of membrane of the cubic nanometre size that we shall be led to consider. Thus the main difference between ρ_1 and ρ_2 in this example comes from the difference in the molecular polarization of the cell wall in a changed electric field. In §8 (proposition 8.9C and equation 8.9) it will be shown, at least for a fairly crude membrane model, that this difference is insufficient at the scale required. It is because of this result that it is necessary to keep our channels localized by requiring frequent “collapse”.

Turn now to the other question; that of whether the quantum state of a channel can with high probability be “collapsed” regularly into neighbourhoods of the states ρ_1 and ρ_2 . We shall attack this question by considering what our observations tell us in various circumstances about the expected (or average) state of a channel. I have claimed, on the one hand, that, after a short period without observation, this expected state will be diffused in space over the membrane, and diffused in time over

the firing cycle, and that it is straightforward to “collapse” out of these diffusions. Now we consider, on the other hand, indirect observations that tend to alter the entire internal structure of the state and that cannot be resolved by simple state decomposition.

We do observe our own brains in more ways than simply by being aware. Awareness provides a mirror of the world, and according to the [sodium channel model](#), that mirror can be constructed from the statuses of sodium channels. We now have to ask to what extent the sodium channels can see themselves in their own mirror. For example, if a man has a fever, then he is aware of his sweats and his chills and his general discomfort. He is not directly aware of his elevated blood temperature. Medical science has discovered, however, that the signs are evidence for the symptom. In other words, we know empirically that our physical constitutions are such that if we feel feverish, then, most likely, we “have a temperature”. A consequence of elevated blood temperature, of course, is a brain and sodium channels which are hotter than usual. Thus the sodium channels of our patient do mirror themselves (whether he knows it or not) as having altered in state.

Mean human blood temperature is around 310°K , with a maximum range from 300°K , at which temperature the oxidation of glucose ceases and a state of suspended animation ensues, to 316°K , when death ensues ([Bell, Emslie-Smith, and Paterson 1980](#), chapter 26). In the next section, we shall model the quantum states of a sodium channel as equilibrium states of a system with N independent degrees of freedom using the specific definition given by hypothesis IV of §5. The larger N is, the more these quantum states are influenced by the ambient temperature. If N is greater than 674, then we can use (8.2) to show that a change in temperature from 300°K to 316°K alters the state by more than the variation allowed by our definition. The point is, that if N is greater than 674, then for $\lambda = \frac{316}{300}$, the right hand side of (8.2) is greater than l , and this violates the constraint imposed by inequality (5.5). As a consequence, we can claim that quantum switches as defined can only be constructed from fairly small portions of body temperature macromolecules. Such a very large temperature change, of course, is very rare, but even if we limited ourselves to a one degree change from 309.5°K to 310.5°K , which is the sort of change in human blood temperature that can be produced by physical exertion and somewhat underestimates the standard diurnal variation, we would not be able to take N to be greater than 175,000. This is of the order of magnitude of the number of degrees of freedom, excited at body temperature, of a complete macromolecule. This analysis of the effect of temperature variation should be enough to demonstrate that the theory being presented is sufficiently specific to be potentially falsifiable by neurophysiological information.

An alternative way of estimating the effect of a change in ambient temperature on a physical system is given by proposition 8.9A. This estimates the effect in terms of specific heat. Again, the larger the system, the greater the influence of temperature change. By this method, the maximum volume of a system that could be a quantum switch is given by a^3 , where estimates of a , in nanometres, range, depending on what the switch is made of, from 1.1 to 1.5, if a temperature change from 300°K to 316°K is allowed, and from 6.8 to 9.7, if a temperature change from 309.5°K to 310.5°K is

allowed. In both cases, the smaller value corresponds to a switch made of water, and the larger to one of phenol, with a variety other organic liquids and solids giving values between these. The values are those which make the right hand side of (8.6) equal $1.1(\text{nm})^3$ will accomodate about 7.5 amino acid residues (an average computed from table 2 of (Chothia 1975), and may be compared with the total volume of the sodium channel protein of about $200(\text{nm})^3$ (as assigned by Begenisich (1987) and Greenblatt, Blatt, and Montal (1985)).

There are other situations in which a subject is aware of changes which may affect the quantum state of his sodium channels without causing him to lose consciousness. These fall into three categories. Firstly, there are large scale environmental effects like changes in ambient pressure, or undergoing an NMR brain scan, or experiencing weightlessness; secondly, there are chemical changes caused by the ingestion of any sort of substance; and thirdly, there are changes in local regions of the brain caused by its own functioning.

As far as changes of the first category are concerned, accepting the nanometre dimension of a switch, just derived from the analysis of natural temperature variations makes the switches too small to be significantly affected by sub-lethal amounts of any such change that I can think of.

The second category is more interesting. A neural membrane is bathed by fluids containing many different chemical species with continually varying concentrations. We can construct a very simple model to measure the effect of such changes in concentration, for species which do not specifically bind to sodium channels, by considering a switch to consist of a small volume (V) of the given species in solution at the physiological concentration. At the nanometre scale, effects are only significant for the most highly concentrated species, which are the various atomic ions. The highest concentrations reached are for sodium and chloride outside the cell and for potassium inside. They are each less than 0.2 mol l^{-1} . The combined solute concentration in blood plasma is around 0.3 mol l^{-1} . Proposition 8.9B below shows that the quantum state of an ideal solution in a fixed volume does not change by more than would be allowed for variations in fixed status switch states as long as the change in number of solute particles in the volume is limited to changes from an initial number N_1 to a final number N_2 which satisfy $N_1^{\frac{1}{2}} \log N_1/N_2 \leq 1$, where, without significant consequence, we assume that $N_1 \geq N_2$.

Given an initial concentration c_1 , the following table gives values of c_1/c_2 sufficient to have $N_1^{\frac{1}{2}} \log N_1/N_2 = 1$, where N_i is the number of particles in volume V at concentration c_i :

| | $V: (10^{-8}\text{m})^3$ | $(2 \times 10^{-9}\text{m})^3$ | $(10^{-9}\text{m})^3$ |
|--------------------------|--------------------------|--------------------------------|-----------------------|
| $c_1:$ | | | |
| 0.3 mol l^{-1} | 1.08 | 2.3 | 10 |
| 0.2 mol l^{-1} | 1.10 | 2.8 | 18 |

Suppose that someone is given a potion by a chemist or by a witch, and that having drunk it, although feeling pretty odd, he is still prepared to claim to be conscious. If we accept that the physical manifestation of the consciousness of the subject can only be a pattern of switching of the kind defined in §5, which uses switches of volume $(10^{-8}\text{m})^3$ and if the effect of the potion was to change the concentration of solutes in those switches by more than ten per cent from 0.2mol l^{-1} , then, we must require that the claim of the subject is false. This is possible, although you might think that that would depend on who the subject is. Normal values of sodium concentrations in cerebrospinal fluid vary from 144mmol l^{-1} to 152mmol l^{-1} (Bell, Emslie-Smith, and Paterson 1980, Table 19.10). This suggests that we might have to face the dilemma just sketched were we indeed to take the volume of a switch to be as large as $(10^{-8}\text{m})^3$. Once again a nanometre scale is indicated.

The above argument is very crude. In particular, in the [sodium channel model](#), we are taking the bulk of a switch to consist of protein, so it is not clear that any of the region that we choose to contain our element need be penetrated by the fluid medium. However, the orders of magnitude involved are interesting, and the sodium gate is certainly in contact with fluid.

In the category of chemical changes, we must also consider those molecules, like some in scorpion toxin, which bind specifically to sodium channels. Such molecules usually have a drastic effect on consciousness, and, in such cases, it is acceptable that the result of binding with such a molecule is to cause loss of awareness of the sodium channel concerned. However, there is also the hypothetical possibility of a paradox, as one can imagine a toxin which could radically change the quantum states of a channel but not disturb its function. In this case, one might have a person whose consciousness had radically altered, while his functioning had hardly changed at all. One might even think of that person as having been, at least in part, turned into a computer. Such potential paradoxes are an inevitable consequence of seeking a definition of consciousness by structure rather than by function. It does not, in fact, seem to me that such paradoxes vitiate the theory, but they should not be ignored.

The third category of situation in which a subject is aware of changes which may affect the quantum states of his sodium channels, concerns changes due to the functioning of the brain itself. Since we are proposing a model in which the awareness of a subject is built from awareness of every neural firing, he will, by definition, be aware, albeit at a sub-verbal level, if a given neuron has fired with high frequency over a long period. This means that we must consider “exhaustion” effects, or, in other words, whether one can tell just from the quantum state of a small portion of a sodium channel that the neuron that it belongs to is being over-worked. The other effect that we might consider in the present category is that of changes in the local electric field at a channel due to the firing patterns of neighbouring neurons, but these changes can be neglected in comparison with the much larger electric field changes that we shall come to.

One possible exhaustion effect might arise from changes in local ionic concentrations due to the repeated firing. Such changes certainly occur. A considerable amount of work has been done on measuring changes in potassium concentrations

outside nerve cells as a result of repetitive firing (see (Kuffler, Nicholls, and Martin 1984, chapter 13) for a review). A resting nerve cell has an external potassium concentration of around 3 mmol l^{-1} . Under artificially repetitive stimulation, this may change to as much as 20 mmol l^{-1} . However, if the size of our quantum switch is taken to be $(2\text{nm})^3$, and the switch is taken to be entirely fluid filled, then applying proposition 8.9B again shows that only a change from 3 mmol l^{-1} to 35 mmol l^{-1} would be sufficient to change the switch state by more than the fluctuations that we would be prepared to permit. This, and similar calculations for the other ionic species, suggest that this sort of concentration change is more relevant because it alters the electric potential across the membrane than because of direct changes in local ionic numbers.

The electric field in which a channel sits is affected, not only by the exhaustion effect, but also by other observable changes. Most neurons in a human brain probably have much more complex electrical behaviours than might be suggested by what has been said so far in this paper. Neurophysiologists have done their most detailed work on the axons of giant neurons from squid. It is now realised that these simple systems do not display the full range of behaviours of mammalian central nervous system neurons. Some of the possible complexities are reviewed by Crill and Schwindt (1983). For example, one class of cerebellar neurons display a substantial modulation of the normal sodium dependent action potential consequent on the opening of calcium channels (Llinás and Sugimori 1980). These calcium channels react directly to the input to the cells, and produce action potentials of their own which have a slower time course than those of the sodium channels.

Such complexities suggest that instead of thinking of a sodium channel as being in a cell which always rests at a potential of, for example, -70 mV , and always fires to a maximum potential of, for example, $+30 \text{ mV}$, we should be prepared, at least, to think of the local resting potential of the cell as varying slowly over a range from, for example, -85 mV to, for example, -50 mV .

Proposition 8.9C and the ensuing derivation of equation 8.9, provide a model according to which this sort of modulation of the background potential of a cell does not have a significant effect on the quantum state of a nanometre dimension volume of sodium channel, except in as far as it alters the probability of the channel being open. In other words, depolarizing a closed channel, even by as much as 100 mV , makes the channel more likely to open, but, if it is seen not to open, then the difference in apparent state due to the change in induced molecular polarization is not significant.

Accepting this argument, means that we can incorporate in our theory, both those neurons which produce graded responses rather than action potentials (Roberts and Bush 1981), and the partial local depolarizations of post-synaptic regions of a neuron which do not necessarily lead to neural firing. The fact that the probability of a channel opening will depend on the ambient electric field is not a problem. There is no reason, under any model, to be concerned if the typical pattern of switching over a neural surface varies with externally observable circumstances; it is only important that the patterns be patterns of true switchings.

The model given by proposition 8.9C is very general, and is, with the caveats mentioned in leading up to equation 8.9, probably applicable. The argument can be

recapitulated by saying that there are parts of the sodium channel which have two metastable states. The occupation probabilities of these states are electric field dependent. Each state will vary continuously with the electric field, but the underlying dichotomy always remains identifiable. It is the continuous variation which is estimated by proposition 8.9C. If we are to satisfy the definition of a switch proposed in §5, then this continuous dependence must not allow either state to change by more than the minimum difference across the dichotomy. According to equation 8.9, this requirement will be satisfied for our nanometre dimension switches. What is essential is that the channel opens with a jump, and that, at any resting potential (resp. firing potential) within the normal range, there is a significant probability of occupying a state which is close to some fixed closed state (resp. some fixed open state). The theory might fail if the channel opened steadily; if, for example, like a voltmeter needle, it moved deterministically, with changing voltage, along a path of rapidly changing states.

Nevertheless, even if this argument is valid, it is conceivable that consideration of variations in electric field might allow us to choose between some of the versions of the sodium channel model. For example, the definition of the states in examples 1, 5, and 6, depend on a clear distinction between resting and firing potentials. Such a distinction may not be possible. In the sliding helices example, the precise state of one particular helix may well depend on the states of the other helices. This could give rise to a problem if the probabilities of occupancy of the states of the other helices vary sufficiently rapidly with voltage. In that case, under a change of voltage, the states of the particular helix that we choose as our switch, may be more affected due to changes in the other helices than they are by the molecular polarization change modelled by proposition 8.9C. Such an effect would require us to adopt a model like that of example 4.

Similar comments can be made about the question of whether a rapidly repeated firing has an observable effect on the state of a sodium channel. Following an action potential, every nerve cell has a “refractory” period during which it is less likely to fire again. This is believed to be caused by the inactivation gates; a neuron cannot fire again until the inactivation gates of an adequate number of its sodium channels have re-opened. If the state of a sliding helix, which forms part of the activation gate, is significantly affected by the status of the inactivation gate, then example 3 would have to be modified; either in the direction of example 4, or, as in example 5, by choosing ρ_1 and ρ_2 to be average states of the helix at appropriate times in the firing cycle.

One of the interesting consequences of the sort of specific definition of consciousness that has been presented in this paper – “consciousness is the existence of a pattern of switching” – is that it is possible to attack from a new angle the old question of whether, complexity aside, a computer or a dog should be thought of as being conscious. In this section, we have been asking the rather more fundamental question of whether, according to our definition, humans are conscious. Needless to say our definition would have to be rejected if we could not give an affirmative answer. I believe that the work of this section has shown that we can give an affirmative answer, but

I have little doubt that further analysis and further neurophysiological information will allow a more precise identification of the class of switches in a human brain.

8. Mathematical Models of Warm Wet Switches.

(for physicists)

In this section, we shall establish the technical estimates used earlier. We shall use (5.5) to test when two quantum states are sufficiently similar to represent one status of a switch. We shall model our switch states as equilibrium density matrices on a Hilbert space \mathcal{H} of the general form $\rho(\beta, \mu, \mathbf{E}) = \exp\{-\beta(H - \mu N - \gamma \mathbf{E} \cdot \mathbf{M})\}/Z$, where, with k Boltzmann's constant, k/β is the temperature, H the Hamiltonian, μ the chemical potential, N the number operator, \mathbf{E} the electric field, \mathbf{M} the electric dipole moment, and γ the local field correction factor. With this model, we assume that our switches interact with the rest of the brain as with a heat bath. We shall assume that H , N , and \mathbf{M} all commute, and, under this assumption, it becomes straightforward to translate the central result below (proposition 8.9) into the mathematics of classical statistical mechanics.

Proposition 8.9 provides simple estimates in terms of thermodynamically defined quantities. It is based mathematically on the inequality (8.3). The preliminary work in this section is aimed at understanding (5.5) and its relation to (8.3). In example 8.3, we compute the norm difference (5.5) for a system of N harmonic oscillators. In proposition 8.4, we prove that, for such a system, the estimate made through inequality (8.3) of this norm difference is greatest in the classical regime. For most of our applications, it is sufficient merely to have the sort of upper bound provided by proposition 8.9, but lemma 8.10 shows that these upper bounds will often be of the correct order of magnitude. Finally, in lemma 8.11, we prove (5.3) for a general von Neumann algebra.

lemma 8.1 For all density matrices σ and ρ on \mathcal{H} ,

$$\|\sigma - \rho\| = 2 \sup\{ |(\sigma - \rho)(P)| : P \text{ is a projection} \}.$$

proof Set $\kappa = \sigma - \rho$.

For any projection P , $2P - 1$ is a unitary map: $(P - (1 - P))^*(P - (1 - P)) = 1$, and so $\|\kappa\| \geq |\kappa(2P - 1)| = |\kappa(2P)|$.

κ can be expressed as a self-adjoint trace class operator (Reed and Simon 1972, §VI.6), so there exists an orthonormal basis $(\psi_n)_{n \geq 1}$ of \mathcal{H} and a sequence $(\lambda_n)_{n \geq 1} \subset \mathbb{R}$, such that $\kappa = \sum_{n=1}^{\infty} \lambda_n |\psi_n\rangle\langle\psi_n|$.

$$\sum_{n=1}^{\infty} \lambda_n = \kappa(1) = \sigma(1) - \rho(1) = 0.$$

Set $P_+ = \sum_{\{n: \lambda_n \geq 0\}} |\psi_n\rangle\langle\psi_n|$, $P_- = \sum_{\{n: \lambda_n < 0\}} |\psi_n\rangle\langle\psi_n|$. P_{\pm} are orthogonal projections with $P_+ + P_- = 1$.

$$\text{Since } \kappa(1) = 0, \kappa(P_+) = -\kappa(P_-) = \sum_{\{n: \lambda_n \geq 0\}} \lambda_n = \frac{1}{2} \sum_{n=1}^{\infty} |\lambda_n|.$$

But, for any bounded operator B , $\kappa(B) = \text{tr}(\kappa B) = \sum_{n=1}^{\infty} \lambda_n \langle\psi_n|B|\psi_n\rangle$, so $|\kappa(B)| \leq \|B\| \sum_{n=1}^{\infty} |\lambda_n|$, and $\|\kappa\| \leq \sum_{n=1}^{\infty} |\lambda_n|$. The result follows. ■

The proof of this lemma shows that

$$\|\sigma - \rho\| = \sum_{n=1}^{\infty} |\lambda_n| = \|\sigma - \rho\|_{\text{tr}}, \quad (8.1)$$

where $\|\sigma - \rho\|_{\text{tr}}$ is the norm of $\sigma - \rho$ considered as a trace class operator.

example 8.2 Suppose $\sigma = |\varphi\rangle\langle\varphi|$, $\rho = |\psi\rangle\langle\psi|$, where $\varphi, \psi \in \mathcal{H}$, with $\|\varphi\| = \|\psi\| = 1$. Then $\|\sigma - \rho\| = 2(1 - |\langle\varphi|\psi\rangle|^2)^{\frac{1}{2}}$. In particular, $\|\sigma - \rho\| = 2$, if and only if $\langle\varphi|\psi\rangle = 0$.

proof Let $\psi = \alpha\varphi + \beta\varphi'$ where $\|\varphi'\| = 1$, $\langle\varphi'|\varphi\rangle = 0$. Then $\sigma - \rho$ has the same eigenvalues as $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} |\alpha|^2 & \alpha\bar{\beta} \\ \bar{\alpha}\beta & |\beta|^2 \end{pmatrix}$.

Using $|\alpha|^2 + |\beta|^2 = 1$, these eigenvalues are

$$\pm|\beta| = \pm(1 - |\alpha|^2)^{\frac{1}{2}} = \pm(1 - |\langle\varphi|\psi\rangle|^2)^{\frac{1}{2}}.$$

The result now follows from equation 8.1. ■

example 8.3 For a set of N independent non-identical harmonic oscillators with common frequency ω ,

$$\|\rho(\beta_1) - \rho(\beta_2)\| = 2 \sum_{n=0}^{N-1} \frac{(N+K)!}{(K+1+n)!(N-1-n)!} \times \\ ((1-t_1)^{N-1-n} t_1^{K+n+1} - (1-t_2)^{N-1-n} t_2^{K+n+1}),$$

where $\rho(\beta)$ is the equilibrium state at temperature k/β , $t_1 = e^{-\beta_1\hbar\omega}$, $t_2 = e^{-\beta_2\hbar\omega}$, we choose $t_1 \geq t_2$, and K is the greatest integer smaller than or equal to

$$\frac{N \log\left(\frac{(1-t_2)}{(1-t_1)}\right)}{\log\left(\frac{t_1}{t_2}\right)}.$$

proof By equation 8.1,

$$\|\rho(\beta_1) - \rho(\beta_2)\| = \sum_{n_1=0}^{\infty} \dots \sum_{n_N=0}^{\infty} |(1-t_1)^N t_1^{n_1+\dots+n_N} - (1-t_2)^N t_2^{n_1+\dots+n_N}| \\ = \sum_{n=0}^{\infty} \frac{(N+n-1)!}{n!(N-1)!} |(1-t_1)^N t_1^n - (1-t_2)^N t_2^n|,$$

where a Taylor expansion of $(1-t)^{-N} = \sum_{n_1=0}^{\infty} \dots \sum_{n_N=0}^{\infty} t^{n_1+\dots+n_N}$ can be used to derive the coefficients.

Using the assumption that $t_1 \geq t_2$, gives

$$\sum_{n=0}^{\infty} \frac{(N+n-1)!}{n!(N-1)!} |(1-t_1)^N t_1^n - (1-t_2)^N t_2^n| \\ = 2 \sum_{n=K+1}^{\infty} \frac{(N+n-1)!}{n!(N-1)!} ((1-t_1)^N t_1^n - (1-t_2)^N t_2^n),$$

since $(1-t_1)^N t_1^n - (1-t_2)^N t_2^n$ is non-positive for $n \leq K$ and non-negative for $n \geq K$, and since

$$\sum_{n=0}^{\infty} \frac{(N+n-1)!}{n!(N-1)!} ((1-t)^N t^n = 1.$$

$$\begin{aligned} \text{Now, } \sum_{n=K+1}^{\infty} \frac{(N+n-1)!}{n!(N-1)!} t^n &= \frac{1}{(N-1)!} \frac{d^{N-1}}{dt^{N-1}} \left(\sum_{n=K+1}^{\infty} t^{N+n-1} \right) \\ &= \frac{1}{(N-1)!} \frac{d^{N-1}}{dt^{N-1}} (t^{N+K} (1-t)^{-1}). \end{aligned}$$

This can be written as a sum over N terms by using the Leibniz formula, and the result follows. ■

This example expresses the norm in a form in which numerical estimation is possible. The norm tends to be larger for low frequencies. In the limit $\omega \rightarrow 0$ (i.e. $\hbar\omega \ll kT$) we have, using Stirling's formula,

$$\|\rho(\beta) - \rho(\lambda\beta)\| \rightarrow 2 \sum_{n=0}^{N-1} \left(\frac{(xN)^n}{n!} e^{-xN} - \frac{(yN)^n}{n!} e^{-yN} \right), \quad (8.2)$$

where $x = \frac{\log \lambda}{\lambda-1}$ and $y = \frac{\lambda \log \lambda}{\lambda-1}$.

In general, it is quite hard to make direct estimates using equation 8.1. An alternative method is provided by the following inequality, proved by Hiai, Ohya, and Tsukada (1981, theorem 3.1) :

For all pairs of density matrices σ and ρ ,

$$\|\sigma - \rho\|^2 \leq 2 \operatorname{tr}(\sigma \log \sigma - \sigma \log \rho). \quad (8.3)$$

proposition 8.4 Choose $\lambda \geq 1$ and let $\rho_{\beta,N}(\boldsymbol{\omega})$ denote the equilibrium state of N independent non-identical harmonic oscillators with frequencies $(\omega_i)_{i=1}^N$ at temperature k/β . Then

$$\|\rho_{\beta,N}(\boldsymbol{\omega}) - \rho_{\lambda\beta,N}(\boldsymbol{\omega})\| \leq \sqrt{2N(\lambda - 1 - \log \lambda)}.$$

proof $\rho_{\beta,N}(\boldsymbol{\omega})$ and $\rho_{\lambda\beta,N}(\boldsymbol{\omega})$ are tensor product states, so,

$$\begin{aligned} \operatorname{tr}(\rho_{\beta,N}(\boldsymbol{\omega}) \log \rho_{\beta,N}(\boldsymbol{\omega}) - \rho_{\beta,N}(\boldsymbol{\omega}) \log \rho_{\lambda\beta,N}(\boldsymbol{\omega})) \\ = \sum_{i=1}^N \operatorname{tr}(\rho_{\beta,1}(\omega_i) \log \rho_{\beta,1}(\omega_i) - \rho_{\beta,1}(\omega_i) \log \rho_{\lambda\beta,1}(\omega_i)). \end{aligned}$$

Using inequality 8.3, it is thus sufficient to prove that, for all $\omega \geq 0$,

$$\operatorname{tr}(\rho_{\beta,1}(\omega) \log \rho_{\beta,1}(\omega) - \rho_{\beta,1}(\omega) \log \rho_{\lambda\beta,1}(\omega)) \leq \lambda - 1 - \log \lambda.$$

Set $t = e^{-\beta\hbar\omega}$. Then

$$\operatorname{tr}(\rho_{\beta,1}(\omega) \log \rho_{\beta,1}(\omega) - \rho_{\beta,1}(\omega) \log \rho_{\lambda\beta,1}(\omega)) = \log \left(\frac{1-t}{1-t^\lambda} \right) + (\lambda-1) \frac{(-t \log t)}{1-t}.$$

$$\text{Set } f_\lambda(t) = \log \left(\frac{1-t}{1-t^\lambda} \right) + (\lambda-1) \frac{(-t \log t)}{1-t}.$$

$\lim_{t \rightarrow 1} f_\lambda(t) = -\log \lambda + \lambda - 1$, so it is sufficient to prove that for all $\lambda \geq 1$, $f_\lambda(t)$ is increasing in t for $t \in [0, 1]$.

$$f'_\lambda(t) = \frac{(\lambda - 1)(1 - t^\lambda)(-\log t) - \lambda(1 - t)(1 - t^{\lambda-1})}{(1 - t)^2(1 - t^\lambda)}.$$

Set $g_\lambda(t) = (\lambda - 1)(1 - t^\lambda)(-\log t) - \lambda(1 - t)(1 - t^{\lambda-1})$. It is sufficient to prove that $g_\lambda(t) \geq 0$ for all $\lambda \geq 1$ and $t \in [0, 1]$.

lemma 8.5 $g_2(t) \geq 0$ for $t \in [0, 1]$.

proof $g_2(t) = (1 - t)(-(1 + t) \log t - 2(1 - t))$.

Set $h(t) = -(1 + t) \log t - 2(1 - t)$. Then $h(1) = 0$ and $h'(t) = t^{-1}(-t \log t + t - 1)$.

Set $u(t) = -t \log t + t - 1$. Then $u(1) = 0$ and $u'(t) = -\log t$.

Thus $u'(t) \geq 0$, $u(t) \leq 0$, and $h(t) \geq 0$. ■

lemma 8.6 $g_\lambda(t) \geq 0$ for $\lambda > 2$ and $t \in [0, 1]$.

proof Using lemma 8.5,

$$\begin{aligned} g_\lambda(t) &\geq 2(\lambda - 1)(1 - t^\lambda)(1 - t)^2 / (1 - t^2) - \lambda(1 - t)(1 - t^{\lambda-1}) \\ &= \left(\frac{1 - t}{1 + t} \right) (\lambda - 2 - \lambda t + \lambda t^{\lambda-1} - (\lambda - 2)t^\lambda). \end{aligned}$$

Set $p_\lambda(t) = \lambda - 2 - \lambda t + \lambda t^{\lambda-1} - (\lambda - 2)t^\lambda$.

$p_\lambda(1) = 0$, $p'_\lambda(t) = -\lambda + \lambda(\lambda - 1)t^{\lambda-2} - \lambda(\lambda - 2)t^{\lambda-1}$,

$p'_\lambda(1) = 0$, $p''_\lambda(t) = \lambda(\lambda - 1)(\lambda - 2)(t^{\lambda-3} - t^{\lambda-2}) \geq 0$,

so $p'_\lambda(t) \leq 0$ and $p_\lambda(t) \geq 0$ for $t \in [0, 1]$. ■

lemma 8.7 $g_\lambda(t) \geq 0$ for $1 < \lambda < 2$ and $t \in [0, 1]$.

proof $g_\lambda(t^{1/(\lambda-1)}) = (\lambda - 1)g_{\lambda/(\lambda-1)}(t)$, so this result can be deduced from lemma 8.6. ■ ■

For N large and $(\lambda - 1)N^{\frac{1}{2}}$ small, the right hand side of 8.2 is asymptotic to $(\lambda - 1)(2N/\pi)^{\frac{1}{2}}$, while $\sqrt{2N(\lambda - 1 - \log \lambda)} = (\lambda - 1)N^{\frac{1}{2}} + O((\lambda - 1)^2)$, so proposition 8.4 and inequality 8.3 can give an estimate of the correct order of magnitude.

lemma 8.8 Let K and L be commuting operators on a finite dimensional Hilbert space. Let $\rho_a = e^{-K-aL}/\text{tr}(e^{-K-aL})$, and write $\langle Y \rangle_a = \text{tr}(\rho_a Y)$. Choose $b \geq a$. Then,

$$\|\rho_a - \rho_b\|^2 \leq (b - a)^2 \sup_{0 \leq x \leq b-a} \left(-\frac{d}{dx} (\langle L \rangle_{a+x}) \right). \quad (8.4)$$

proof This is a consequence of perturbation theory applied to (8.3). Set

$$f(x) = \text{tr}(\rho_a \log \rho_a - \rho_a \log \rho_{a+x})$$

$$= \langle -K - aL - \log(\text{tr}(e^{-K-aL})) \rangle_a + K + (a + x)L + \log(\text{tr}(e^{-K-(a+x)L})) \rangle_a$$

$$= x \langle L \rangle_a + \log \langle e^{-xL} \rangle_a.$$

$$f'(x) = \langle L \rangle_a + \langle (-L)e^{-xL} \rangle_a / \langle e^{-xL} \rangle_a = \langle L \rangle_a - \langle L \rangle_{a+x}.$$

$f(0) = f'(0) = 0$, so Taylor's formula with remainder shows that $f(b-a) = \frac{1}{2}(b-a)^2 f''(x)$ for some $x \in [0, b-a]$.

8.4 now follows directly from 8.3. ■

The restriction to finite dimensional \mathcal{H} is for simplicity of exposition. Note, that in the notation of lemma 8.8,

$$\begin{aligned} f''(x) &= (\langle e^{-xL} \rangle_a \langle L^2 e^{-xL} \rangle_a - \langle (-L) e^{-xL} \rangle_a \langle (-L) e^{-xL} \rangle_a) / (\langle e^{-xL} \rangle_a)^2 \\ &= \langle L^2 \rangle_{a+x} - (\langle L \rangle_{a+x})^2 = \langle (L - \langle L \rangle_{a+x})^2 \rangle_{a+x}, \end{aligned}$$

from which it follows that

$$\begin{aligned} \|\rho_a - \rho_b\|^2 &\leq (b-a)^2 \sup_{0 \leq x \leq b-a} \left(-\frac{d}{dx} \langle L \rangle_{a+x} \right) \\ &= (b-a)^2 \sup_{0 \leq x \leq b-a} \langle (L - \langle L \rangle_{a+x})^2 \rangle_{a+x}. \end{aligned} \quad (8.5)$$

This makes manifest the positivity of the bound (8.4). The relation between (8.4), or (8.5), and thermodynamic quantities is well known, and yields our main result:

proposition 8.9

A) Let $\rho(\beta) = e^{-\beta H} / \text{tr}(e^{-\beta H})$. Then

$$\|\rho(\beta) - \rho(\lambda\beta)\| \leq |\lambda - 1| (c_v a^3 / k)^{\frac{1}{2}} \quad (8.6)$$

where c_v is the specific heat per unit volume, and a^3 is the volume of substance.

B) Let $\rho(\beta, \mu) = e^{-\beta(H - \mu N)} / \text{tr}(e^{-\beta(H - \mu N)})$. Then

$$\|\rho(\beta, \mu_1) - \rho(\beta, \mu_2)\| \leq |\mu_1 - \mu_2| \sup \left(\beta \frac{\partial N}{\partial \mu} \right)^{\frac{1}{2}}, \quad (8.7)$$

and, in particular, for an ideal gas in the classical regime and for a dilute solution,

$$\|\rho(\beta, \mu_1) - \rho(\beta, \mu_2)\| \leq \max\{\sqrt{N_1}, \sqrt{N_2}\} |\log N_1 / N_2|, \quad (8.8)$$

where N_i is the mean number of particles in state $\rho(\beta, \mu_i)$.

C) Let $\rho(\beta, E) = e^{-\beta(H - \gamma \mathbf{E} \cdot \mathbf{M})} / \text{tr}(e^{-\beta(H - \gamma \mathbf{E} \cdot \mathbf{M})})$. Then

$$\|\rho(\beta, E_1) - \rho(\beta, E_2)\| \leq |E_1 - E_2| (\beta \gamma \varepsilon_0 \chi V)^{\frac{1}{2}}$$

where χ is the electric susceptibility and V the volume of the system.

proof

A) Apply 8.4, setting $K = 0$, $L = H$, $a = \beta$, $b = \lambda\beta$.

$$-\frac{d}{d\beta} \langle H \rangle = kT^2 \frac{d}{dT} \langle H \rangle = kT^2 c_v a^3.$$

B) 8.7 is another direct application of 8.4. 8.8 follows, because, for an ideal gas in the classical regime $N = (V/V_Q) e^{\mu\beta}$ where V_Q is the quantum volume, while, for a dilute solution, $\mu = \mu_0 + kT \log c/c_0$, where c is the concentration and the suffix 0 refers to some standard state. (Our applications of this result will be so crude, that there will be no need to distinguish between concentration and activity.)

C) This also follows from 8.4, using $\langle \mathbf{M} \rangle_{\mathbf{E}} = \varepsilon_0 \chi V \mathbf{E}$. ■

The application of C requires some care. There is considerable literature relevant to the task of estimating the electric field acting at a neural membrane (e.g. (Nelson,

Colonomos, and McQuarrie 1975) and (McLaughlin 1977)). There are also various models, depending on circumstances, for estimating γ , but it can be assumed to be of order unity (see Chełkowski 1980). For a crude estimate, we put $\gamma = 1$, $T = 310^\circ\text{K}$, $|\mathbf{E}_1 - \mathbf{E}_2| = 3 \times 10^7 \text{Vm}^{-1}$ - corresponding to a potential drop of 100 mV across a 3 nm membrane, and $V = (1.5 \text{ nm})^3$. Then corresponding to a volume V of membrane of susceptibility $\chi = 2$, we have

$$|\mathbf{E}_1 - \mathbf{E}_2|(\beta\gamma\varepsilon_0\chi V)^{\frac{1}{2}} = 0.11, \quad (8.9)$$

while, even for water of susceptibility $\chi = 80$, we only have

$$|\mathbf{E}_1 - \mathbf{E}_2|(\beta\gamma\varepsilon_0\chi V)^{\frac{1}{2}} = 0.71.$$

If we had 1 instead of 0.11 here, then we could argue that a small patch of membrane would by itself act as a switch, even without the presence of a channel. On the other hand, when using channels, the smallness of 0.11 allows us, as argued in §7, to avoid difficulties with varying background potentials.

Lemma 8.8 and proposition 8.9 provide upper bounds for the norm differences in which we are interested. For most of the applications made in this paper, such upper bounds are sufficient. Thus, in §7, it was enough to know that a quantum switch could be constructed from $1(\text{nm})^3$ of sodium channel; to know that it might be possible to use a volume greater than $1(\text{nm})^3$, was of lesser importance. Even so, it is useful to have an idea of how accurate the approximation provided by lemma 8.8 might be. In combination with (8.5), the next result shows that, if the right hand side of (8.4) equals 1, then $1 \geq \|\rho_a - \rho_b\| \geq \frac{1}{4} + O((b-a)^3)$. It follows that we can expect the estimate provided by proposition 8.9 to be of the correct order of magnitude.

lemma 8.10 *With the notation of lemma 8.8, set $Z(a) = \text{tr}(e^{-K-aL})$. Then*

$$\|\rho_a - \rho_b\| \geq 2 \left(1 - \frac{Z(\frac{1}{2}(a+b))}{\sqrt{Z(a)Z(b)}} \right) = \frac{1}{4}(b-a)^2 \langle (L - \langle L \rangle_a)^2 \rangle_a + O((b-a)^3).$$

proof K and L have common eigenvector expansions, say, $K = \sum_{n=1}^N \kappa_n |\psi_n\rangle\langle\psi_n|$ and $L = \sum_{n=1}^N \lambda_n |\psi_n\rangle\langle\psi_n|$. Then, by (8.1),

$$\|\rho_a - \rho_b\| = \sum_{n=1}^N |e^{-\kappa_n - a\lambda_n} / Z(a) - e^{-\kappa_n - b\lambda_n} / Z(b)|.$$

For $x \geq y \geq 0$, $|x - y| = x - y = (\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y}) \geq (\sqrt{x} - \sqrt{y})(\sqrt{x} - \sqrt{y})$,
so

$$\|\rho_a - \rho_b\| \geq \sum_{n=1}^N \left(\frac{e^{-\frac{1}{2}\kappa_n - \frac{1}{2}a\lambda_n}}{\sqrt{Z(a)}} - \frac{e^{-\frac{1}{2}\kappa_n - \frac{1}{2}b\lambda_n}}{\sqrt{Z(b)}} \right)^2 = 2 \left(1 - \frac{Z(\frac{1}{2}(a+b))}{\sqrt{Z(a)Z(b)}} \right).$$

$$\text{Let } g(x) = \frac{Z(a + \frac{1}{2}x)}{\sqrt{Z(a)Z(b)}}.$$

Then $g(0) = 1$, $g'(x) = \frac{1}{2}g(x)(\langle L \rangle_{a+x} - \langle L \rangle_{a+\frac{1}{2}x})$, $g'(0) = 0$,

$$g''(x) = -\frac{1}{4}g(x)(2\langle (L - \langle L \rangle_{a+x})^2 \rangle_{a+x} - \langle (L - \langle L \rangle_{a+\frac{1}{2}x})^2 \rangle_{a+\frac{1}{2}x} - (\langle L \rangle_{a+x} - \langle L \rangle_{a+\frac{1}{2}x})^2),$$

and $g''(0) = -\frac{1}{4}\langle(L - \langle L \rangle_a)^2 \rangle_a$, so the given approximation is a Taylor expansion. ■

lemma 8.11 For all normal states σ and ρ on a von Neumann algebra \mathcal{A} ,

$$\|\sigma - \rho\| = 2 \sup\{ |(\sigma - \rho)(P)| : P \in \mathcal{A} \text{ is a projection} \}.$$

proof Set $\kappa = \sigma - \rho$.

As in lemma 8.1, for any projection P , $|2\kappa(P)| \leq \|\kappa\|$.

Let $\kappa = \kappa_+ - \kappa_-$ be the decomposition of κ into positive and negative parts and let P_\pm be their support projections; these are proved to exist in most textbooks on von Neumann algebras, e.g. (Strătilă and Zsidó 1979, theorem 5.17).

$$\begin{aligned} \|\kappa\| &\leq \|\kappa_+\| + \|\kappa_-\| = \kappa_+(P_+) + \kappa_-(P_-) \\ &= \kappa(P_+) - \kappa(P_-) = \kappa(P_+) - \kappa(1 - P_+) = 2\kappa(P_+). \end{aligned}$$

9. Towards a More Complete Theory.

“What is an observed phenomenon?” That is the central question for the interpretation of quantum mechanics. Bohr claimed that an observed phenomenon was something that could be described in classical terms, while von Neumann claimed that it was an eigenvector of a self-adjoint operator. In this paper it has been proposed that an observed phenomenon is a pattern of switching in a human brain. Many questions remain to be tackled before this proposal can be confirmed as the foundation of a complete interpretation of quantum mechanics. These include:

- 1) How does a pattern of switching correspond to an awareness?
- 2) Is the definition of a priori probability, given by (5.6), satisfactory?

While the first question can be considered from many angles, it is essentially a re-statement of the age-old mind-body problem, and cannot be “solved” in any absolute sense. As a re-statement it is interesting because it involves a definition of “body” which is much more precise than is usual. Indeed, one goal of this paper is to emphasize that, in a quantum mechanical universe, “body”, in this sense, is the ultimate observed phenomenon, and, as such, the manner of its existence cannot be taken for granted.

The second question is of a more physical character. Some of the questions it leads to are:

- 3) What can be said about the state ω , defined in §5 as the state of the universe prior to any “collapse”?
- 4) What are the mathematical properties of the function $\text{app}_{\mathcal{B}}$?
- 5) In what sense does the world external to the switches in an observer’s brain exist in a state approximating the state in which he might describe it as appearing to exist?
- 6) Does the predicted a priori probability of appearance of a given experimental result agree with the observed frequency of such results?
- 7) Are the apparent observations of different observers compatible?

For several years now, I have been working on the analysis of this theory. My belief is that the questions above can be dealt with, and, in particular, that questions

six and seven can be answered affirmatively. I hope eventually to publish a book on the topic. At present, a preliminary draft of this book exists. The purpose of this paper has been to abstract from the book an aspect of the theory which should be of independent interest.

Acknowledgements. I am grateful to A.M. Donald and S.F. Edwards for useful conversations and comments on the manuscript, and to the Leverhulme Trust for financial support.

References.

- Aldrich, R.W., Corey, D.P., and Stevens, C.F. 1983 A reinterpretation of mammalian sodium channel gating based on single channel recording. *Nature* **306**, 436–441.
- Angelides, K.J., Elmer, L.W., Loftus, D., and Elson, E. 1988 Distribution and lateral mobility of voltage-dependent sodium channels in neurons. *J. Cell Biol.* **106**, 1911–1925.
- Armstrong, C.M. 1981 Sodium channels and gating currents. *Physiol. Rev.* **61**, 644–683.
- Begenisich, T. 1987 Molecular properties of ion permeation through sodium channels. *Ann. Rev. Biophys. Biophys. Chem.* **16**, 247–263.
- Bell, G.H., Emslie-Smith, D., and Paterson, C.R. 1980 *Textbook of Physiology*, 10th ed. Churchill-Livingstone: Edinburgh.
- Burns, B.D. 1968 *The Uncertain Nervous System*. Edward Arnold: London.
- Cantrell, C.D. and Scully, M.O. 1978 The EPR paradox revisited. *Physics Reports* **43**, 499–508.
- Catterall, W.A. 1986a Molecular properties of voltage-sensitive sodium channels. *Ann. Rev. Biochem.* **55**, 953–985.
- Catterall, W.A. 1986b Voltage-dependent gating of sodium channels – correlating structure and function. *Trends NeuroSci.* **9**, 7–10.
- Chełkowski, A. 1980 *Dielectric Physics*. Elsevier: Amsterdam.
- Chothia, C. 1975 Structural invariants in protein folding. *Nature* **254**, 304–308.
- Crill, W.E. and Schwindt, P.C. 1983 Active currents in mammalian central neurons. *Trends NeuroSci.* **6**, 236–240.
- Davies, E.B. 1974 Markovian master equations. *Commun. Math. Phys.* **39**, 91–110.
- Dieudonné, J. 1969 *Foundations of Modern Analysis*. Academic Press: New York.
- Dodge, F.A. 1961 Ionic permeability changes underlying nerve excitation. In *Biophysics of Physiological and Pharmacological Actions* (ed. A.M. Shanes), pp. 119–143. American Association for the Advancement of Science: Washington D.C.

- Donald, M.J. 1986 On the relative entropy. *Commun. Math. Phys.* **105**, 13–34.
- Donald, M.J. 1987a Further results on the relative entropy. *Math. Proc. Camb. Phil. Soc.* **101**, 363–373.
- Donald, M.J. 1987b Free energy and the relative entropy. *J. Stat. Phys.* **49**, 81–87.
- Driessler, W., Summers, S.J., and Wichmann, E.H. 1986 On the connection between quantum fields and von Neumann algebras of local operators. *Commun. Math. Phys.* **105**, 49–84.
- Eccles, J.C. 1973 *The Understanding of the Brain*. McGraw-Hill: New York.
- Eccles, J.C. 1986 Do mental events cause neural events analogously to the probability fields of quantum mechanics? *Proc. R. Soc. Lond.* **B 227**, 411–428.
- Exner, P. 1985 *Open Quantum Systems and Feynman Integrals*. Reidel: Dordrecht.
- French, R.J. and Horn, R. 1983 Sodium channel gating – models, mimics, and modifiers. *Ann. Rev. Biophys. Bioeng.* **12**, 319–356.
- Greenblatt, R.E., Blatt, Y., and Montal, M. 1985 The structure of the voltage-sensitive sodium channel. *FEBS Lett.* **193**, 125–134.
- Guy, H.R. and Seetharamulu, P. 1986 Molecular model of the action potential sodium channel. *Proc. Natl. Acad. Sci. (U.S.)* **83**, 508–512.
- Haag, R. and Kastler, D. 1964 An algebraic approach to quantum field theory. *J. Math. Phys.* **5**, 848–861.
- Hiai, F., Ohya, M., and Tsukada, M. 1981 Sufficiency, KMS condition, and relative entropy in von Neumann algebras. *Pac. J. Math.* **96**, 99–109.
- Hille, B. 1984 *Ionic Channels of Excitable Membranes*. Sinauer: Sunderland, Massachusetts.
- Hodgkin, A.L. and Huxley, A.F. 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol. (Lond.)* **117**, 500–544.
- Hofstadter, D.R. 1979 *Gödel, Escher, Bach: an eternal golden braid*. Harvester Press: Sussex.
- Houslay, M.D. and Stanley, K.K. 1982 *Dynamics of Biological Membranes*. Wiley: Chichester.
- Korn, H. and Faber, D.S. 1987 Regulation and significance of probabilistic release mechanisms at central synapses. In *Synaptic Function* (ed. G.M. Edelman, W.E. Gall, & W.M. Cowan), pp. 57–108. Wiley: New York.
- Kosower, E.M. 1985 A structural and dynamic molecular model for the sodium channel of *Electrophorus electricus*. *FEBS Lett.* **182**, 234–242.
- Kubo, R., Toda, M., and Hashitsume, N. 1985 *Statistical Physics II*. Springer-Verlag: Berlin.
- Kuffler, S.W., Nicholls, J.G., and Martin, A.R. 1984 *From neuron to brain*, 2nd ed., Sinauer, Sunderland Massachusetts.
- Llinás, R. and Sugimori, M. 1980 Electrophysiological properties of *in vitro* Purkinje cell somata and dendrites in mammalian cerebellar slices. *J. Physiol. (Lond.)*, **305**, 171–195 & 197–213.
- London, F.W. and Bauer, E. 1939 *La Théorie de l’Observation en Mécanique Quantique*. Hermann et Cie.: Paris.

- Mackey, G.W. 1963 *The Mathematical Foundations of Quantum Mechanics*. Benjamin: New York.
- McLaughlin, S. 1977 Electrostatic potentials at membrane-solution interfaces. *Curr. Top. Membr. Transp.* **9**, 71–144.
- Mermin, N.D. 1985 Is the moon there when nobody looks? Reality and the quantum theory. *Physics Today* (April), 38–47.
- Nelson, A.P., Colonomos, P. and McQuarrie, D.A. 1975 Electrostatic coupling across a membrane with titratable surface groups. *J. Theor. Biol.* **50**, 317–325.
- von Neumann, J. 1932 *Mathematische Grundlagen der Quantenmechanik*. Springer-Verlag: Berlin.
- Noda, M., Shimizu, S., Tanabe, T., Takai, T., Kayano, T., Ikeda, T., Takahashi, H., Nakayama, H., Kanaoka, Y., Minamino, N., Kangawa, K., Matsuo, H., Raftery, M.A., Hirose, T., Inayama, S., Hayashida, H., Miyata, T., and Numa, S. 1984 Primary structure of *Electrophorus electricus* sodium channel deduced from cDNA sequence. *Nature* **312**, 121–127.
- Noda, M., Ikeda, T., Kayano, T., Suzuki, H., Takeshima, H., Kurasaki, M., Takahashi, H., and Numa, S. 1986a Existence of distinct sodium channel messenger RNAs in rat brain. *Nature* **320**, 188–192.
- Noda, M., Ikeda, T., Suzuki, H., Takeshima, H., Takahashi, T., Kuno, M., and Numa, S. 1986b Expression of functional sodium channels from cloned cDNA. *Nature* **322**, 826–828.
- Reed, M. and Simon, B. 1972 *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press: New York.
- Roberts, A. and Bush, B.M.H. (Eds.) 1981 *Neurons without Impulses*. Cambridge University Press: Cambridge.
- Salkoff, L., Butler, A., Wei, A., Scavarda, N., Giffen, K., Ifune, C., Goodman, R., and Mandel, G. 1987 Genomic organization and deduced amino acid sequence of a putative sodium channel gene in *Drosophila*. *Science* **237**, 744–749.
- Scott, A.C. 1977 *Neurophysics*. Wiley: New York.
- Srinivasan, Y., Elmer, L., Davis, J., Bennett, V., and Angelides, K. 1988 Ankyrin and spectrin associate with voltage-dependent sodium channels in brain. *Nature* **333**, 177–180.
- Strătilă, S. and Zsidó, L. 1979 *Lectures on von Neumann Algebras*. Abacus Press: Tunbridge Wells.
- Wheeler, J.A. and Zurek, W.H. 1983 *Quantum Theory and Measurement*. Princeton University Press: Princeton.
- Wigner, E.P. 1961 Remarks on the mind-body question. In *The Scientist Speculates* (ed. I.J. Good), pp. 284–302 Heinemann: London.