Applying mathematical theories of consciousness to quantum collapse theories

Kelvin J. McQueen & David J. Chalmers

December 15, 2016

Contents

1 Introduction

A familiar story about quantum mechanics runs as follows. Quantum-mechanical systems are describable by a wavefunction. Most of the time, the wavefunction evolves according to the deterministic Schrodinger equation. The wavefunction need not specify definite properties for the system: instead it may specify a superposition of many different values for position, momentum, and other properties. When one measures these properties, however, one always obtains a definite result. After measurement, the system's wavefunction is now in a new state that specifies this definite value. The result of the measurement and the resulting wavefunction are determined probabilistically by the pre-measurement wavefunction of the system according to the Born rule, which associates wavefunction amplitudes with probabilities.

The canonical version of this story was given by John von Neumann in *Mathematical Foundations of Quantum Mechanics* (1932). Construed as an empirical apparatus for predicting the results of measurements, this story has been tremendously successful. The predictions made by the story have been borne out again and again, and it has been used to explain all sorts of phenomena. As a result, the empirical apparatus has long ago obtained the status of orthodoxy.

Because of this empirical success, it is natural to construe the story as a description of the reality underlying quantum mechanics. Taken at face value, the story suggests that quantum-mechanical reality fundamentally involves a wavefunction with a bipartite dynamics. First, there is the Schrodinger evolution, which is linear, deterministic, and constantly ongoing. Second, there is a process of collapse into a definite state, which is nonlinear, nondeterministic, and happens only on certain occasions of measurement.

This story about quantum-mechanical reality has met with much less widespread acceptance than the corresponding story about empirical predictions. One problem is that the process of collapse is somewhat mysterious and quite unlike any other process in physics. The biggest problem, though, is what has come to be known as the measurement problem (see Albert 1992; Bell; Wallace 2008). The story contains a fundamental principle saying that collapses happen when and only when a measurement occurs. But on the face of it, the notion of "measurement" is vague and anthropocentric, and is inappropriate to play a role in a fundamental specification of reality. At the very least, one needs a much more specific proposal about how measurement is to be understood and about how it could play a role in fundamental physics. No such proposal has attracted much in the way of support.

Because of this, physicists and philosophers interested in the foundations of quantum mechanics have largely turned away from this face-value interpretation of quantum mechanics and moved toward various other interpretations. Perhaps the closest to the face-value interpretation are spontaneous collapse interpretations (Ghirardi et al, Pearle), which still give a fundamental role to the collapse process, but which hold that collapses occur randomly and not as the result of a measurement process. Another class of interpretations holds that collapses are not fundamental: instead the effects of collapse can be derived from the Schrodinger equation alone, which predicts that quantum superpositions decohere as they interact with macroscopic systems in their environment. There are also interpretations that do without collapse altogether. These include many-worlds interpretations (Everett), on which the wavefunction never collapses and instead evolves into a superposition of many branches even at the macroscopic scale, and hidden-variables interpretations (Bohm), on which the wavefunction serves to guide a separate layer of quasi-classical particles with definite positions and other properties.

Still, we think that the potential of interpretations of quantum mechanics in the mold of the face-value interpretation has not yet been adequately explored. There is a class of precise and rigorous interpretations of quantum mechanics that more closely resemble the face-value interpretation than any of the interpretations above, without giving a fundamental role to the imprecise notion of measurement. This is the class of *M-property theories*. According to these theories, there is a fundamental process of collapse which occurs whenever a certain sort of triggering event obtains. (These contrast with spontaneous collapse interpretations, according to which collapses occur randomly.) On the traditional interpretation, the triggering event is measurement. But to solve the measurement problem, we have to replace measurement by some more precise class of triggering events. A wide class of possible triggering events is possible, yielding a wide class of M-property theories. It is this class of interpretations that we explore in what follows.

A specific M-property theory that we will explore in the final section is one on which the property (the *M-property*) that triggers the collapse is consciousness. The idea that consciousness collapses the wave function has a long history in quantum mechanics (London and Bauer, Wigner, Stapp) but has often been exaggerated and ridiculed. We think that it is possible to make an M-property theory along these lines precise and well-motivated. At the same time, the class of M-property theories goes well beyond this sort of interpretation, and there are many versions that give no special role to consciousness at all. We will start by developing a general model for M-property theories, and then look at various hypotheses about the nature of the M-properties.

We stress that we do not know whether an M-property theory is correct. We are exploring such an interpretation rather than endorsing it. In particular, we are not asserting that these interpretations are superior to other interpretations of quantum mechanics. Both of us have some sympathy with many-world interpretations and think that hidden-variable and spontaneous-collapse interpretations cannot easily be excluded. But we think that M-property theories deserve close attention. If it turns out that they have fatal flaws, they can be set aside. But if they have no clear fatal flaws, then they should be taken seriously as possible descriptions of quantum-mechanical reality.

2 Taking textbook quantum mechanics at face value

One of the goals of M-property theory is to stay as close as possible to textbook quantum mechanics. There is good reason for this: the textbook theory is the orthodox formulation of quantum theory that has been used to advance quantum theory over the last century, so we should not abandon its basic structure without good reason. And so to begin with, the goal is to set the measurement problem aside for one moment and consider what the world must be like if the textbook theory is true. However, as soon as we do so, we immediately run into an ambiguity: what is more fundamental to the theory, a measurement *property*, or a measurement *process*?

On the first disambiguation, the theory describes a special class of objects (the measuring devices) which each possess a special property (a measurement property) which is responsible for causing collapse. On the second disambiguation, the theory describes a special kinds of process (the measurement process) which is responsible for collapse. The first disambiguation provides the conceptual foundations for m-property theory. The second disambiguation provides an alternative framework which already starts to look highly complex, and threatens to bring in quasi-mental notions right from the start. However, is one that has been explored in detail by Henry Stapp [2011 [etc.]] and will be worth considering before developing M-property theory.

To understand Stapp's theory it is useful to isolate the components of his theory that are additions to the textbook theory. Stapp [2011: 24] postulates four processes, processes 1 to 4. Process 3 corresponds to the collapse postulate of textbook quantum mechanics. Process 2 corresponds to the Schroedinger equation of textbook quantum mechanics. Process 1 corresponds to the measurement process, and is meant to explain process 2. Process 1 corresponds to an experimenter posing a question to nature (e.g. is the particle spin-up or spin-down?). The idea is that nature responds to the posing of such question by collapsing the wave-function and yielding a definite answer. We then (finally) have process 0 which is defined as “some process that is not described by quantum theory, but determines the [process 1] ‘free choice”’.

The problems with Stapp’s account are clear, and begin with process 0. Process 0 is *by stipulation* not describable by physics, which leaves us with no precise account of process 0, including no account of why or when it happens. But if we have no account of why or when process 0 happens, then we have no account of why or when process 1 happens. But if we have no account of why or when process 1 happens, then we have no account of why or when process 3 happens. But then we have no account of why or when collapse occurs, meaning we have no solution to the measurement problem. Let us therefore consider M-property theory instead.

At the basis of M-property theory is the idea that there is a special class of objects that each possess a special property responsible for collapse - a measurement property. Ultimately, we want to remove reference to the imprecise notion of ‘measurement’ and so instead of speaking of measurement properties we shall speak of M-properties, which is a stand-in for some precise property responsible for collapse. To develop M-property theory, we must first clarify the idea of being *responsible* for collapse.

On the most straightforward reading of this idea, an M-property is a property that is governed by a basic law which states that the property *refuses* superposition and responds to impending superpositions by wave-function collapse. We can model the idea as follows. We might imagine a particle (particle “p”) that begins in some definite location X and then evolves, in accord with the Schroedinger equation, into a superposition of two different positions, which we can label H (for ‘here’) and T (for ‘there’). (Perhaps the particle was sent towards a well calibrated beam-splitter which put the particle into a superpostion of being transmitted and being deflected, but the details won’t matter.) We can represent this deterministic Schroedinger evolution as follows:

$$|X\rangle_p \rightarrow \alpha\,|H\rangle_p + \beta\,|T\rangle_p \tag{1}$$

Now imagine we have some measuring device (device “d”). This device is designed so that it will display on its screen “H” if the particle is found *Here* and “T” if the particle is found *There*. In the meantime its screen simply displays “R” for “ready to measure”. Schroedinger evolution entails that when this device interacts with the particle, the particle’s superposition will be magnified up into the device, causing the device to enter into a superposition of displaying “H” and displaying “T”:

$$(\alpha\,|H\rangle_p + \beta\,|T\rangle_p)\,|\text{“}R\text{”}\rangle_d \rightarrow \alpha\,|H\rangle_p\,|\text{“}H\text{”}\rangle_d + \beta\,|T\rangle_p)\,|\text{“}T\text{”}\rangle_d \tag{2}$$

In the textbook theory, the interaction that led to the device superposition was a *measurement*. The device-particle system therefore stops evolving in accord with the deterministic Schroedinger equation and instead evolves in accord with the collapse postulate. Accordingly, the system collapses in one of two possible ways:

$$\alpha\,|H\rangle_p\,|\text{“}H\text{”}\rangle_d + \beta\,|T\rangle_p)\,|\text{“}T\text{”}\rangle_d \rightarrow |H\rangle_p\,|\text{“}H\text{”}\rangle_d \tag{3}$$

Or:

$$\alpha\,|H\rangle_p\,|\text{“}H\text{”}\rangle_d + \beta\,|T\rangle_p)\,|\text{“}T\text{”}\rangle_d \rightarrow |T\rangle_p\,|\text{“}T\text{”}\rangle_d \tag{4}$$

The probability of evolution (3) is given by the absolute value square of α, the probability of evolution (4) is given by the absolute value square of β. This is the Born rule.

Now, M-property theory does not say that collapse happens because a measurement occurs. Rather, collapse happens because an M-property (whatever that property may be) was put into superposition. To illustrate, assume that the device happens to possess an M-property. Moreover, the device's M-property is correlated with device's screen as follows: when the screen reads "H" the device has M-property M_1, and when the screen reads "T" the device has M-property M_2, and when the screen reads "R" the device has M-property M_0. Then, when the particle interacts with the device, the Schroedinger equation entails:

$$(\alpha \left|H\right\rangle_p + \beta \left|T\right\rangle_p) \left|``R''/M_0\right\rangle_d \rightarrow \alpha \left|H\right\rangle_p \left|``H''/M_1\right\rangle_d + \beta \left|T\right\rangle_p) \left|``T''/M_2\right\rangle_d \tag{5}$$

We may no say that collapse is not caused by measurement, but is caused by the fact that an M-property was put into superposition. The M-property, given the fundamental law governing its behaviour, yields either:

$$\alpha \left|H\right\rangle_p \left|``H''/M_1\right\rangle_d + \beta \left|T\right\rangle_p) \left|``T''/M_2\right\rangle_d \rightarrow \left|H\right\rangle_p \left|``H''/M_1\right\rangle_d \tag{6}$$

Or:

$$\alpha \left|H\right\rangle_p \left|``H''/M_1\right\rangle_d + \beta \left|T\right\rangle_p) \left|``T''/M_2\right\rangle_d \rightarrow \left|T\right\rangle_p \left|``T''/M_2\right\rangle_d \tag{7}$$

with probabilities given by the Born rule.

There are two aspects of the theory most in need of clarification. Firstly, are there any candidate precisely definable properties that could be M-properties? Secondly, can we be more precise about how the basic law governing M-properties works? Let us consider each question in turn.

3 M-properties

Different hypotheses for different M-properties yield empirically distinct M-property theories. Therefore, to constrain the possible class of candidate M-properties we should first consider empirical constraints. The two most basic empirical constraints come from (i) empirical data involving quantum effects such as interference effects, and (ii) the empirical fact that our measurements always have definite outcomes. For example, the first constraint rules out an interpretation on which the m-property is position: there are many experimental results, such as results of double-slit experiments, showing that wavefunctions are not always in an eigenstate of position. The second constraint appears to rule out an interpretation on which the m-property is the property of being a unicorn, or the property of being a rare isotope: on these interpretations, measurement devices and human perceptual processes would never collapse, and neither measurement nor perception would yield determinate results.

For an M-property theory to satisfy these constraints, the corresponding m-properties must be such that (i) we have not observed quantum effects indicating superpositions of m-properties, and (ii) m-properties are present in measurement processes or at least in human perception, and they covary with the observed results of measurements.

The second constraint ensures that m-properties will be entangled with measurement results so that if the former have a definite value, so will the latter. In effect, the first constraint requires that variable m-properties are not too ubiquitous (if even electrons have variable m-properties, then we have probably observed superpositions of m-properties already), while the second constraint requires that they are not too rare.

The first constraint entails that m-properties cannot be familiar fundamental physical properties such as position, mass, charge, and spin. For all of these, superpositions have been demonstrated. So if m-properties are physical properties, they must be nonfundamental properties. It is natural to suppose that fundamental physical entities will have no nontrivial m-properties, and that m-properties obtain at a somewhat higher level.

A natural structure for an m-property, then, is one such that fundamental physical entities have a null value for that quantity, and likewise for other very small physical systems. But beyond a certain threshold (of complexity or size or some other scale), systems have variable and non-null m-properties. This variation helps make it the case that a non-superposition constraint on m-properties can help to collapse wave functions. (Consider as an analogy: electrons do not display meter readings, while measuring devices have variable meter readings.)

Some potential m-properties include:

(1) molecular energy properties: the energy of a molecule (if greater than a certain threshold, else zero).

(2) configurational properties: say, the energy of a system consisting of a number of particles involved in a certain pyramidal configuration (if the system has that configuration, else zero).

(3) Spacetime structure: spacetime cannot be in a superposition of two different curvatures [Penrose, Diosi].

(4) informational properties: for example, Tononi's property *phi*, which measures the amount of information integration in a system (if greater than a certain threshold, else zero).

(5) mental properties, such as a system's state of consciousness (if it is conscious, else zero).

Versions of all of these properties plausibly satisfy constraint (i) above: at least if we choose appropriate thresholds, no quantum-mechanical experiments to date have indicated superpositions of these properties. As for constraint (ii), mental properties satisfy this constraint by definition, and the right choice of informational property will almost certainly satisfy the constraint as well. It is not obvious whether the mass properties and the configurational properties above satisfy the constraint, but it is not implausible that a judiciously chosen version of these properties will satisfy the constraint.

For the reasons discussed earlier, none of these properties are fundamental physical properties. Nevertheless all of them are precise and well-defined, or if not, there are certainly precise and well-defined properties in the vicinity. More generally, we can certainly expect that there will be a number of precise and well-defined triggered-collapse interpretations that meet these minimal empirical constraints.

One might object that there are too many potential candidates for m-properties that meet the relevant empirical constraints, and that consequently the form of a triggered-collapse interpretation is underdetermined. If this is the only objection to triggered-collapse theories, it is hard to see that it is a serious objection to their truth: at worst we have an embarrassment of riches, with a multitude of adequate interpretations. In any case, empirical methods can in principle distinguish among these interpretations, as we discuss in the next section.

4 The basic law governing M-properties

Our initial analysis of an M-property was that it was the property responsible for the collapse of the wavefunction. We then stated that on the most straightforward reading of this idea, an M-property is a property that is governed by a basic law which states that the property *refuses* superposition and responds to impending superpositions by wave-function collapse. But this idea cannot be correct. For no property cannot *absolutely* refuse superposition due to the quantum Zeno effect. In what follows we explain this idea. The solution is then to clarify the basic law governing m-properties so that m-property superpositions are unstable.

4.1 The quantum Zeno effect

Roughly, the qauntum Zeno effect states that frequent quantum measurement makes it hard for measured quantities to change. This effect causes fatal problems for the idea that m-properties absolutely refuse superpositions.

Here is the basic idea. For any given physical property, if a system evolves from having one value v_1 of that property, to another v_2, it must go through superpositions of v_1 and v_2, such that the probability of the initial value v_1 continuously decreases from one. Consequently, if the M-property collapses immediately upon superposition, it will immediately collapse to its initial value. The M-property then gets stuck!

This basic idea deserves further elaboration. Why must it be the case that to get from v_1 to v_2, the system must go through superpositions of those values? There are two cases to consider. The first case involves pure Schroedinger evolution applied to a closed system. Here the important point is that the states v_1 and v_2 are represented by *orthogonal* vectors in the system's vector space. Schroedinger evolution continuously rotates the state vector and cannot make 90 degree jumps. It is for this reason that Schroedinger evolution gaurantees that evolution from v_1 to v_2 must go through superpositions. The second case involves Schroedinger evolution applied to an open system interacting with an external system. The interesting case is where the external system (through entanglement) imposes a superposition on the system. It may at first seem that entering into this superposition does not require that the system's initial state continuously decrease from probability one. But in fact it does, since the system enters into a superposition of interacting with the external system and not interacting with the external system. Not interacting with the external system is the initial state whose probability continuously descends from zero. And so, if the first superposition causes collapse, collapse will favour the initial state, and the systems cannot even interact.

Should we therefore abandon m-property theory as well as the consciousness causes collapse hypothesis? We think not. We think the quantum Zeno effect should be viewed as another tool for refining m-property theory. Thus, we should not understand the basic law governing m-properties as stating that m-properties absolutely refuse superposition. Instead, the basic law states that m-property superpositions are *unstable*. As we shall see, there are a number of ways of making this idea precise based on the zoo of dynamical collapse theories that have been developed over the past few decades. In what follows we consider what we take to be the most plausible formulations of the basic law of m-properties.

4.2 Spacetime curvature as the m-property

The theory of Diosi [1987] and Penrose [2014] is an example of an m-property theory, where the m-property is *spacetime curvature*. On their theory, it is of the very nature of spacetime itself that spacetime curvature refuses superposition and responds with collapse. In particular, if we have a superposition of two spacetime curvatures, then the greater the difference in those curvatures, the more unstable the superposition, and the fast it will collapse.

Penrose [2014] formulates a precise description of this instability by defining the survival time of the superposition as $t = \hbar/\mathrm{E}_G$. The quantity E_G is the gravitational self-energy of the difference between the mass distributions of the two states in the superposition. This is an instructive model for our purposes. The basic structure involves defining a distance measure among values of the chosen m-property (in this case, spacetime curvature). This will also be the basic structure of our first proposal in section 4.4.1 below.

4.3 Entangled particle number as the m-property

It is an interesting conceptual question to ask whether the GRW-Pearle model is an example of an m-property theory. According to the (conceptually simpler) GRW model, every particle has a probability per unit time for spontaneous collapse. The collapse involves the particle's position wave-function being multiplied by a Gaussian collapse function. The point in the position wave-function that the Gaussian collapse function is centred on is chosen by the Born rule. The GRW theory therefore introduces two new constants, the probability per unit time for spontaneous collapse, and the width of the Gaussian.

Is the GRW theory an example of an m-property theory? On the face of it, it might seem so. The more entangled particles within a given system, the higher the probability per unit time that the system will collapse to a definite position. Consequently, it might seem like entangled particle number is being treated as an m-property. On the other hand, the motivation behind the GRW theory was to move away from the idea that some high-level property is responsible for collapse, and let collapse just be an aspect of the underlying dynamics. In particular, the idea was to just make it seem as if large objects were responsible for collapse, when really, this appearance receives a deeper explanation in terms of stochastic equations of motion. For this reason, we think the GRW model (and Pearle model) should not be understood within the m-property framework, not least because "being made of many particles" is a strange property to endow causal powers to. Either way, the GRW model offers an illuminating model that inspires our second proposal in section 3.2: just as GRW models collapse frequency as a function of the amount of particles in a system, we can model collapse frequency as a function of the amount of consciousness in a system, as measured by the integrated information theory of consciousness.

4.4 Integrated information theory

To make the consciousness causes collapse hypothesis precise we consider neuroscientific proposals for the physical correlate of consciousness and whether they can be treated as proposing m-properties. Our first two hypotheses come from considering the integrated information theory (IIT) of consciousness.

4.4.1 MICS as the m-property

The first hypothesis uses the property that IIT treats as the physical correlate to a system's consciousness: the system's maximally irreducible conceptual structure (MICS).

According to the integrated information theory, the amount of consciousness in a system is the amount of integrated information it has, and the quality of experience is specified by the informational relationships it generates. Qualia space (Q) is a space having an axis for each possible state (activity pattern) of a complex. Within Q, each submechanism specifies a point corresponding to a repertoire of system states. Arrows between repertoires in Q define informational relationships. Together, these arrows specify a quale — a shape that completely and univocally characterizes the quality of a conscious experience. This shape is defined as the system's maximally irreducible conceptual structure (MICS). (Phi, the quantity of consciousness associated with the experience, corresponds to the height of this shape.) See Balduzzi and Tononi [2009].

The fundamental law describing how a system's MICS can act as its m-property is best formulated analogously to the Penrose-Diosi model of collapse. On their view, spacetime curvature is the m-property, and as the difference in the curvature between two components of a curvature superposition rises, the probability of collapse to one of those curvatures rises. Analogously, we can say that as the difference in the MICS between two components of a MICS superposition rises, the probability of collapse to one of those MICSs rises. The difference between two MICSs can arguably be defined in terms of distance in qualia space. A key part of the project involves making this hypothesis precise.

4.4.2 Phi as the m-property

The second hypothesis uses the property that IIT treats as the physical correlate to a system's *amount* of consciousness: the system's amount of integrated information (phi). The fundamental law describing how a system's phi can act as its m-property is best formulated analogously to the GRW-Pearle model of collapse. On their view, the probability per unit time for spontaneous collapse (to position) is a function of the system's particle number. Analogously, we can say that the probability per unit time for spontaneous collapse (either to phi or to position) is a function of the system's phi.

Recently, a version of this hypothesis has been explored by Kobi Kremnizer and André Ranchin [2015]. On their model, the probability per unit time for spontaneous collapse (to position) is a function of the system's *quantum* integrated information. That is, they formulate a quantum notion of integrated information for a system in terms of the reduced density matrices of the components of a that system, and let the system's collapse rate be a function of that value. Their proposal has some advantages. But one disadvantage is that it is not clear how their notion of quantum integrated information relates to Tononi's notion of integrated information. Consequently, it is not clear how their notion relates to consciousness. We therefore propose to develop this idea in terms of Tononi's classical notion of integrated information. But in the process, we will compare and contrast with the alternative theory of Kremnizer and Ranchin.

4.5 Experimental tests

Conventional tests of dynamical collapse theories involve diffraction experiments, and experiments aimed at testing the predicted amount of energy conservation violation. Such tests can be used to decide between m-property hypotheses. Consequently, they can be adapted to directly test the consciousness causes collapse hypotheses we have considered.

Diffraction experiments set out to observe interference effects. The famous example is the double slit experiment. This experiment has been conducted with objects the size of buckyballs. The size of such objects are therefore not m-properties. But what if we were to somehow endow buckyballs (or something similar) with varying amounts of the m-property? The M-property theories currently under consideration predict that the more *phi* such objects have, the less probability they will exhibit interference effects. Perhaps we can program nanocomputers not much larger than buckballs with varying levels of *phi*. If so, we can test these theories.

Roger Penrose has proposed an experiment to test his spacetime causes collapse theory, called FELIX (free-orbit experiment with laser interferometry X-rays), in which an X-ray laser in space is directed toward a tiny mirror, and fissioned by a beam splitter from tens of thousands of miles away, with which the photons are directed toward other mirrors and reflected back. One photon will strike the tiny mirror moving en route to another mirror and move the tiny mirror back as it returns, and according to conventional quantum theories, the tiny mirror can exist in superposition for a significant period of time. This would prevent any photons from reaching the detector. If Penrose's hypothesis is correct, the mirror's superposition will collapse to one location in about a second, allowing half the photons to reach the detector. One could imagine in principle that the mirror could somehow be endowed with integrated information to affect these results in an observable way. (Perhaps we just need to strap one of Aaronason's compact disks to the mirror.)

Because position basis collapses tend to add energy to the system, the system's temperature increases. On IIT M-property theories, as the system possesses more *phi*, the probability that the system's particle's undergo GRW-like collapses rises. That way we can make use of the GRW formulas for calculating collapse induced temperature increase. To test these theories, then, we simply need to look for more frequent temperature variation in systems with varying levels of *phi*.

For the same reason that collapses produce spontaneous heating, they also produce spontaneous emissions of sound. The amounts of sound generated by a system, on the current m-property theory, will be a function of the GRW calculations of sound emission of the collapse of a single particle, together with the amount of particles in the system and how much of the m-property the system has. This means that may mean that ultra-sound experiments on sleeping and waking brains could be potentially be used to test the IIT m-property theories.

4.6 Mathematical modeling

The Schroedinger equation may be stated as:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H, \rho(t)] \tag{8}$$

The equation of motion for the reduced density matrix of an open system may be represented by the Lindblad equation:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H, \rho(t)] + \sum_{n,m=1}^{N^2-1} h_{n,m}(L_n\rho(t)L_M^\dagger - \frac{1}{2}(\rho(t))L_M^\dagger L_n + L_M^\dagger L_n\rho(t))) \tag{9}$$

The terms additional to the Schroedinger equation describe the effects of the environment of the system. These terms may also be used to describe the effects of the collapse of the wave function. Kremnizer and Ranchin [2015] realised that a theory which lets collapse rate be a function of phi can be modeled accordingly:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H, \rho(t)] + \sum_{n,m=1}^{N^2-1} h_{n,m}(\phi(\rho(t)))(L_n\rho(t)L_M^\dagger - \frac{1}{2}(\rho(t))L_M^\dagger L_n + L_M^\dagger L_n\rho(t))) \tag{10}$$

Here, h$_n, m$ are Hermitian matrix elements that are continuous functions of the integrated information of ρ (all zero when phi(ρ(t)) = 0). L$_k$ is a basis of operators on the N-dimensional system Hilbert space, which determine the collapse basis.

We aim to model all the various hypotheses we are developing by similar master equations.

5 Questions and Objections

5.1 What about the conservation of energy?

If m-properties involve position, collapse onto eigenstate yields a big violation of conservation of energy. We could try postulating a GRW-style collapse that multiplies the wavefunction by a Gaussian instead. Or better, we can appeal to m-properties other than position. For example, m-properties involving energy will not violate conservation of energy.

5.2 What about the tails problem?

One won't have tails for m-properties, but one will for other properties. For example, if energy is an m-property, then position will involve superpositions with infinite tails. That's OK – as long as this predicts the results of measurements. Do objects have locations on this picture? The spatial functionalism outlined in Chalmers (2012) offers a way to vindicate the claim that they do.

5.3 What about the quantum Zeno effect?

Worry: in certain circumstances, continuous measurement means that a measured value can never change. This applies especially to discrete quantities – so perhaps m-properties had better be continuous. Even here, there is a question of how m-properties can ever evolve out of their default "zero" state (how consciousness could first evolve, for example!). The worry is that the moment that the amplitude of nonzero states creeps above zero, it will immediately collapse back to zero. Figure this out!

5.4 Can M-property theories be reconciled with relativity?

Presumably m-properties will be relativistically invariant. But collapse doesn't seem to be. It happens at a time – in which reference frame? There have been attempts to make spontaneous collapse models relativistically invariant – look at these. There's also been some work on relativistic stochastic Schrodinger equations.

5.5 Nonfundamental m-properties can't enter into fundamental laws

It seems unusual for nonfundamental properties to enter into the fundamental laws of collapse. But it doesn't seem bizarre or incoherent. Really what one has is an arbitrary operator that plays a special role in the fundamental laws, by constraining the wavefunction to always remain in an eigenstate of that operator. One might balk at the arbitrariness – why this operator and not that one? But arbitrary operators seem no worse off than arbitrary constants, however. One can raise the same question – why this value for the constant and not that one? But such constants seem to enter into the fundamental laws nevertheless.

Furthermore, there is one remaining hypothesis on which m-properties are fundamental.

6 Consciousness as the M-property

The idea that consciousness collapses the wave function has a long history. Von Neumann (1932) hints at it, London and Bauer (1939) make the idea explicit, and Wigner (1961) has a well-known informal discussion of the idea. The idea has been prominent in some popular treatments of quantum mechanics, such as Zukav (xx) and Capra (1975). But there has been surprisingly little work on developing a detailed theory along these lines. The most notable recent development of such a theory is by Henry Stapp, who as we discussed earlier, pursues an avenue quite different from the one we pursue here.

By 'consciousness', what is meant is phenomenal consciousness, or subjective experience. A system is conscious when there is something it is like to be that system, from the inside. A mental state is conscious when there is something it is like to be in that state. Conscious states come in many flavors and varieties. Perhaps the most obvious conscious states are ordinary perceptual states: there is something it is like to see colors and shapes, and indeed to perceive pointer locations. It is natural to think of perception as involving a causal chain from objects to the environment to the eye and then to the brain, culminating in a conscious perceptual experience.

The view that consciousness collapses the wave function can be specified in the current framework as a triggered collapse intepretation in which the m-property is consciousness. This m-property will take a null value when a system is unconscious. When the system is conscious at a certain time, the system's m-property will be precisely the state of consciousness that it is in at that time: that is, the total conscious experience of that system at that time. Given that consciousness is an m-property, systems can never be in superpositions of two different states of consciousness.

To illustrate the view, we can suppose that there are systematic correlations between certain central brain processes and consciousness. Suppose an electron in a superposition —a¿ + —b¿ registers on a measurement device and then the result is perceived by a human subject. Assuming the measurement device is not conscious, than at the first stage the electron and the device will go into an entangled state of $|a > |M(a) > +|b > |M(b) >$. Once the result reaches the brain,then at least setting aside the constraint above, we would expect the electron, device, and brain will go into an entangled state $|a > |M(a) > |B(a) > +|b > |M(b) > |B(b) >$. But the brain states correlate with consciousness (not much changes if one takes the two to be identical), so this would yield an entangled superposition $|a > |M(a) > |B(a) > |C(a) > +|b > |M(b) > |B(b) > |C(b) >$. But consciousness cannot be superposed, so the system will collapse into $|a > |M(a) > |B(a) > |C(a) >$ or $|b > |M(b) > |B(b) > |C(b) >$, with Born-rule probabilities deriving from the operator associated with consciousness. In effect, just at the point where the measurement is reaching consciousness, the electron, the measurement device, and the brain will collapse into a definite state.

Why think that the m-property is consciousness, as opposed to any other property? There are perhaps five main motivations (ordered roughly in terms of increasing strength): conceptual, epistemological, explanatory, metaphysical, and causal.

The first motivation comes from the conceptual connection between consciousness and measurement. It is arguable that the core pretheoretical idea of measurement is that of measurement by a conscious observer. If this is right, the standard hypothesis of collapse on measurement leads to a consciousness-collapse view. One could respond that the pretheoretical notion of measurement is looser than this. But even if so, the view will at least provide a relatively precise and nonarbitrary way to clarify the imprecise concept of measurement and the imprecise claim that collapse happens on measurement. Other clarifications are certainly possible, as we have seen earlier, but all seem to involve a degree of arbitrariness. It is also arguable that consciousness is a precise, non-vague property on which we have a clear pretheoretical grasp. To take the m-property to be consciousness itself provides a nonarbitrary theory that fits well with the standard form of the collapse framework.

A second (and related) motivation is epistemological. The consciousness-collapse view is especially well-suited to save what is arguably the central ”determinate measurement” datum: that we never consciously experience superposed states. On the current view, such superposed experiences are automatically ruled out. On any other view, the connection with the datum will be more indirect. Most m-properties will not guarantee the truth of the datum: one can find cases where consciousness and m-properties are dissociated, so that nonsuperposition of m-properties (along with other laws) will not entail nonsuperposition of consciousness. There may be some special m-properties (in particular, those tied to the physical preconditions for consciousness in brains and related systems) that cannot be dissociated in this way and that therefore support the entailment. As a result, this motivation (like the first) does not provide a knockdown argument for the consciousness-collapse view. Still, the view provides an especially neat and tight way of saving the datum.

A third motivation is explanatory: the view arguably provides a sort of explanation of why the collapse constraint is true. It is arguable that it follows from the nature of consciousness that consciousness cannot be superposed. For consciousness to be superposed, there would have to be superposed total states of consciousness: for example, a subject who is in a superposition of a total state involving an experience of redness at a location and a different total state involving experience of blueness at that location. It is arguable that there is no way to make sense of this suggestion. The best we can do is imagine two different subjects of consciousness, or a subject with a sort of complex two-field state of consciousness, or subjects to whom objects seem to be both blue and red. But none of these would really be a superposition of total states: the first would be two separate total states, and the second and third would involve a single complex total state.

[[[Wigner (1961) seems to appeal to something like this motivation when he suggests that the hypothesis that a conscious being is in a superposed state "appears absurd because it implies that my friend was in a state of suspended animation".]]]

If this is right, then superposed states of consciousness are not just unfamiliar: they are inconceivable and perhaps metaphysically impossible. This marks a difference between consciousness and position, energy, and the like, where superpositions are not so hard to grasp. The key difference is perhaps that we have some direct acquaintance with the nature of consciousness, which seems to rule out superpositions. It must be admitted the issues are somewhat murky here, and perhaps something could be said to defend superposed states of consciousness. But if it is correct that the nature of consciousness rules out superposition, then this would provide a distinctive explanation of why the collapse law is true.

The fourth motivation is metaphysical. On one philosophical view of consciousness, property dualism, consciousness is a fundamental nonphysical property of reality, not reducible to or explainable in terms of fundamental phyical properties such as spacetime, mass, and charge. If one accepts this view, there will be distinctive motivations for a consciousness-collapse view. For a start, it will yield a view on which the m-properties that bring about collapse are fundamental (if nonphysical) properties, so that the fundamental collapse law involves only fundamental properties. It also yields an attractive view where purely physical dynamics are always governed by the Schrodinger equation. Collapses only come about due to the intervention of an extra-physical element, namely consciousness. If one already has reason to believe in this extra-physical element, then the hypothesis that it brings about collapse leads to an especially elegant picture of the world.

The final motivation is causal. Many have raised questions about the causal role of consciousness in a physical world. These questions are especially pressing for the dualist, but they also arise for the physicalist. No-one has a clear idea of exactly what consciousness does. The consciousness-collapse view provides a clear answer to that question by giving a causal role to consciousness. Consciousness is what triggers wave-function collapse. It is not hard to extend this role to a causal role for consciousness in governing behavior, as I explain shortly. So if one takes it to be a pretheoretical datum that consciousness plays a causal role, consciousness-collapse interpretations can vindicate that datum.

The fourth and fifth motivations raise the issue of physicalism and dualism. It should be noted that the consciousness-collapse view is quite compatible with both physicalism and dualism. One can consistently hold that consciousness is a physical property, and that physical property is the m-property that triggers wave function collapse. One can also consistently hold that consciousness is a nonphysical property, and that this nonphysical property is the m-property that triggers wave function collapse. The physicalist view has the advantage that there is no need to postulate extra ontology, and that m-properties can be represented in straightforward physical terms. The dualist version has the advantage that m-properties are fundamental, and that purely physical dynamics are uniformly governed by the Schrodinger equation.

Consciousness-collapse interpretations are often rejected precisely because they are associated with dualism. The point above suggests that this association is not cut-and-dried. But at the same time, *if* one has independent reason to accept dualism about consciousness, this gives reason to take these consciousness-collapse interpretations very seriously. The fourth and fifth motivations above carry special weight here: consciousness-collapse interpretations allow a fundamental trigger for collapse, and they give a fundamental causal role to consciousness.

Our view is that there are serious philosophical reasons to accept a property-dualist view of consciousness on which consciousness is a fundamental property. This is not the place to elaborate those reasons in depth, but one key idea is that physical processes only explain the structure and dynamics of complex systems, and that more than this is required to explain consciousness. Physical structure and dynamics suffices to explain the "easy problems" of explaining cognitive functions and behavior, but not the "hard problem" of why all this structure and dynamics is associated with conscious experience. This suggests that consciousness cannot be explained in terms of the existing fundamental properties of physics: spacetime, mass, and so on. If this is right, then science requires that we expand the catalog of fundamental properties. Taking consciousness itself to be a fundamental property is the natural result.

Of course such a view is highly controversial, not just among physicists but among philosophers. But it is worth noting that the central reason that most philosophers give to reject property dualism is the problem of mental causation: how could nonphysical mental properties play a causal role in the physical world? The current picture gives a quite coherent picture on which consciousness plays such a causal role: it plays the key causal role of triggering wave function collapse.

In fact, the standard philosophical argument against dualism is an argument from physics: (1) mental properties affect physical properties, (2) physics is causally closed, in that every property that affects a physical property is a physical property, so (3) mental properties are physical properties. The argument for (2) is that physics leaves no "gaps" where mental properties could do causal work. But in fact, our leading current physical theories leave room for a large such gap, precisely at the point of wave function collapse. (One might even suggest that had a deity wanted to design physical laws that leave room for consciousness, she could not have done much better than this.) So the argument from physics carries little weight in the current context.

Instead, we are left in the odd situation wherein philosophers reject property dualism by an appeal to physics (physics is causally closed), while physicists reject consciousness-collapse interpretations for broadly philosophical reasons (the interpretations are dualistic). It is clear that taken together, these reasons to reject dualist consciousness-collapse interpretations do not have much force. Perhaps there are other reasons to reject consciousness-collapse interpretations, or other reasons to reject dualism, but these familiar reasons on their own cannot do the work.

We think that the dualist consciousness-collapse view should be taken seriously. So it is worth spelling out the view a little more, and addressing some questions and objections.

The best way to think about the dualist consciousness-collapse view is as follows. Purely physical dynamics is governed by the Schrodinger equation and other laws of physics. These laws are supplemented by *psychophysical* laws connecting physics to consciousness in both directions. In the physics-to-consciousness direction, we have laws specifying that certain sorts of physical properties are associated with certain sorts of consciousness. To oversimplify, we can suppose the law says that some complex physical property P is associated with consciousness (and that different values of P are associated with different conscious states). In the consciousness-to-physics direction, we have the collapse law, which specifies how impending superpositions of consciousness resolve probabilistically into a definite state of consciousness and an associated wave function collapse.

This view immediately faces any number of questions:

6.1 How can states of consciousness be represented in the wave function?

This is no problem for the physicalist version, on which consciousness is a physical property representable just like any other physical property. By contrast we are not used to representing nonphysical properties in wave functions. Here there are two choices. First, we can extend the the formalism so that states of consciousness are included in the underlying space that yields the configuration space within which the wave function sits. Second, we can leave the wave function as purely physical, and still invoke the Born rule whenever the wavefunction is about to enter a superposition of physical states each of which corresponds to a different state of consciousness.

6.2 Consciousness is still redundant?

Someone might object that the view still leaves consciousness causally redundant. On a dualist consciousness-collapse interpretation, there will typically be a physical property P that correlates perfectly with consciousness. One can then develop a *physicalist* collapse interpretation on which collapse is brought about by this physical property P, and not by consciousness. There will at least be a possible world (we might think of it as a quantum zombie world) where collapse works this way. In this world, the physical wave function will evolve just as in our world. So consciousness may seem redundant.

In response: on the dualist interpretation, it will be consciousness that directly causes the wave function to collapse, with the physical property P only indirectly causing the collapse by first causing the mediating conscious state. So consciousness is causally relevant to physical processes here. Furthermore, if one accepts the third motivation above on which the nature of conscousness *explains* wave function collapse (collapse is brought about by consciousness in virtue of its nature), then one will have a key explanatory role for consciousness in behavior as well. To be sure, a quantum zombie world may still be possible, but it will be a world in which wave function collapse is less well-explained than it is in our world.

One might also worry: in the actual world, how do we know that it is consciousness that triggers collapse, and not property P? I think the answer here is that either hypothesis is available, but insofar as we already have reason to believe that consciousness is a fundamental property, then the hypothesis that consciousness triggers collapse is a much simpler and more attractive one. The hypothesis has at least three advantages. First, this way the fundamental law of collapse involves a fundamental property. Second, we have a better explanation of collapse, along the lines above. Third, this way we have a causal role for consciousness, cohering with a strong pretheoretical intuition. These virtues of simplicity, explanatory power, and coherence all give reasons to favor the view over the alternative.

6.3 Consciousness plays the wrong sort of causal role?

One might also worry that consciousness-collapse interpretations do not give consciousness the *kind* of causal role that we pretheoretically would expect it to have. There are at least two worries here, both stemming from the fact that we expect consciousness to produces distinctive effects of behavior. Pretheoretically, we expect consciousness to bring about large qualitative differences in behavior. We expect it to be responsible for most intelligent behavior, and certainly for some intelligent behavior such as actions that follow conscious decisions, and verbal reports such as utterances of 'I am conscious'.

One worry is that the most obvious effects of collapse point the wrong way: collapse of consciousness will collapse perceived objects such as measurement instruments, but what we want is for consciousness to affect action. In response, we can note that a collapse of consciousness will collapse an associated brain state, and this brain state will be entangled with action states or will at least cause a corresponding action state, so a collapse of consciousness will help bring about a determinate action. For example, if consciousness probabilistically collapses into an experience of red rather than an experience of blue, this collapse will bring about a corresponding state in the perceptual areas of the brain, which may itself lead to an utterance of 'I am experiencing red' rather than 'I am experiencing blue'.

It is also worth noting that consciousness is not just limited to perceptual experience. There is also agentive experience, the experience of agency and action: say, the experience of choosing to lift one's left hand rather than one's right hand. We can imagine that even after perceptual experience collapses brain states associated with perception, the brain will sometimes evolve into superposed brain states associated with agency, leading to potential superpositions of agentive experience. If consciousness is an m-property, one course of agentive experience (the experience of choosing to lift one's left hand) will be selected. As a result, the brain will collapse into the corresponding physical state, and typically a corresponding course of action (lifting one's left hand) will also be selected. So one's agentive experience will play a clear causal role in action.

This picture naturally raises issues about free will. On this view, the experience of choice plays a nondeterministic causal role in bringing about action. On some popular conceptions of "free will", on which what matters for free will is nondeterminism and a role for consciousness, this picture may vindicate free will in the relevant sense. Others may object that the choices themselves are themselves selected probabilistically, and that random choices are no better than deterministic choices when it comes to free will. We think the issues are far from straightforward, so we will set aside issues about free will here, but we note that a causal role for consciousness can be expected to have some bearing on those issues.

This leads to the second worry: that if collapses due to consciousness accord with the Born rule governing probabilities, then consciousness at best plays a sort of dice-rolling role. It will probabilistically select between different available outcomes, but it will not give us a qualitatively different outcome. After all, under a hypothesis where physical property P collapsed the wave function, purely physical quantum zombies would have behaved the same way. So consciousness will not make outcomes on which humans behave intelligently or on which they say 'I am conscious' any more likely than they would have been if some other property had collapsed the wave function. One might even simulate the dynamics in a classical computer (with a pseudorandom number generator), with no role for consciousness, and the same patterns of behavior would ensue.

In response, we are inclined to concede that most of what this objector says is correct. The quantum zombie scenario suggests that there is a sort of structural/mathematical explanation that might be given for our actions without mentioning consciousness. Still, this structural explanation would not provide a *complete* explanation of our actions, precisely because it leaves out the role of consciousness in grounding that structure. (Like many structural explanations, it leaves out the actual causes.) In the actual world consciousness is causing the relevant behavior, and consciousness may explain why it is that we behave determinately at all. One might have liked a stronger, more transformative causal role for consciousness that could not even in principle have been duplicated without conscousness, but it is not clear why such a role is essential.

If one does want a stronger role for consciousness, the most obvious move is to suggest that the role for consciousness in collapse is not entirely constrained by the Born probabilities. Perhaps perceptual consciousness obeys those constraints (thereby explaining our observations in quantum experiments), but agentive experience does not. For example, collapses due to agentive experience might be biased in such a way that more ”intelligent” choices that lead to more intelligent behavior tend to be favored than they would be according to the Born rule. This picture sacrifices the great simplicity of the original quantum dynamics, and it could perhaps be disconfirmed through the right sort of experiments and simulations, but it is arguable that our current evidence leaves room open for it. We do not find this picture especially attractive, but it is at least worth putting it onto the table.

6.4 Non-fundamental properties in fundamental laws?

An opponent might object that even on a consciousness-collapse view, there will need to be fundamental psychophysical laws connecting property P (the physical correlate of consciousness) to consciousness. Furthermore, P cannot be a fundamental physical property: if it were, we would be left with a panpsychist collapse view on which superpositions would not persist long enough to generate the familiar quantum-mechanical results. So we still have nonfundamental properties involved in fundamental laws. In response, one can concede the basic point, while noting that it is a problem already faced by any dualistic approach to consciousness (panpsychism aside). If we are already dualists, then the consciousness-collapse view will at least restrict the role of nonfundamental properties will be restricted to the psychophysical law governing the distribution of consciousness, and will leave them out of the laws governing physical dynamics. So compared to other collapse interpretations, the consciousness-collapse view at least minimizes the role of nonfundamental properties in fundamental laws.