Consciousness and the Collapse of the Wave Function

David J. Chalmers and Kelvin J. McQueen

Draft 18/04/2015

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 Schrödinger 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 Schrödinger 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; Maudlin 1995; 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), 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,[1] 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 *triggered collapse* interpretations of quantum mechanics.

A triggered collapse interpretation is one according to which 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 triggered collapse interpretations. It is this class of interpretations that we explore in what follows.

A specific triggered-collapse interpretation that we will explore in the final section is one on which the triggering event involves consciousness. The idea that consciousness collapses the wavefunction 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 a triggered-collapse interpretation along these lines precise and well-motivated. At the same time, the class of triggered-collapse interpretations goes well beyond this sort of interpretation, and there are many versions of a triggered-collapse interpretation that give no special role to consciousness at all. We will start by developing a general model for triggered-collapse interpretations, and then look at various hypotheses about the nature of the triggers.

We stress that we do not know whether a triggered-collapse interpretation 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-

[1] **DAVE: how are the last two views distinct?**

collapse interpretations cannot easily be excluded. But we think that triggered-collapse interpretations 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.

Triggered Collapse Interpretations

The familiar collapse-upon-measurement dynamics for quantum mechanics can be put as follows.

1. Schrödinger equation: When no measurement occurs the wavefunction evolves deterministically and linearly.

2. Collapse postulate: When a measurement occurs, the wavefunction collapses into an eigenstate of the operator associated with the measured quantity q, with probabilities given by the Born rule.

To convert these dynamics into a precise interpretation, we have to replace the vague notion of "measurement" by something more precise. A first idea is to replace "measurement" in "a measurement occurs" by some more precisely defined class of event: "an m-event occurs", perhaps, where the class of m-events is precisely defined. However, it is not quite as simple as this. The dynamics do not just mention the class of measurement events; they also make essential reference to the notion of measuring a quantity. So to make these dynamics precise, we would have to introduce a precisely defined relation to quantities, which we might call the m-relation. Such a picture already starts to look highly complex, and threatens to bring in quasi-mental notions right from the start.

Fortunately there is a simpler way to proceed. The key is to start from a version of the traditional story that appeals to the notion of a measurement device. Measurement devices are a privileged class of entities, privileged by the fact that they never enter into superpositions.[2] More precisely, there will be a privileged class of *measurement properties*, which are instantiated by only some entities (the measurement devices). We can think of measurement properties intuitively as akin to meter readings or pointer locations.

On this picture, there is a fundamental principle entailing that measurement properties never enter into superpositions. Suppose that we have a quantum system in a superposed state $a|A> +b|B>$, and the system interacts with a measurement device in such a way that, if it were not for this principle, would yield an entangled superposition $a|A>|M(A)> +b|B>|M(b)>$, where $|M(A)>$ and $|M(b)>$ are the measurement properties brought about by $|A>$ and $|B>$. Then because of the principle, the system will instead evolve into a collapsed state $|A>$

[2] As we shall see, this constraint can be slightly weakened, e.g. allowing for superpositions of nearly identical states or allowing for superpositions of low values of the m-quantity.

$|M(A)>$ or $|B>|M(B)>$, with probabilities given by the Born rule associated with the measurement property.

On this way of doing things, there is no essential appeal to the notion of measuring arbitrary quantities. Instead, we appeal only to the notion of a measurement property. This still leaves the interpretation imprecise, but we can now easily generalize it to a class of precise interpretations. We need only replace the appeal to measurement properties by an appeal to an arbitrary precise property, which we might call an m-property.[3] For any such m-property, there will be a corresponding triggered-collapse interpretation of quantum mechanics, deriving from the constraint that m-properties never enter into superpositions.

Equivalently, one can appeal to the m-operator corresponding to the m-property, and impose the constraint that the wavefunction of a system must always be in an eigenstate of the m-operator.[4] We can then construe the collapse postulate informally as follows: when the Schrödinger equation (taken alone) predicts that a system is about to evolve into a non-eigenstate of the m-operator, the system instead evolves into a corresponding eigenstate of the m-operator, with probabilities given by the Born rule.

This collapse postulate may sound loose, but it is not hard to make it precise. However, the most obvious way of making it precise - which retains the dynamics of ordinary quantum mechanics - runs into immediate difficulties. By imposing the strict stipulation that the system must always be in an eigenstate of the m-operator, we have the dynamics of continuous measurement (Jacobs and Steck 2006). That is, we have the dynamics that would obtain (on a traditional measurement interpretation) if the m-property were being *continuously measured* by an outside observer.

If (it's as if) the m-property is continuously measured like this, then it will simply freeze. This is the *quantum Zeno effect*, and is a quite general consequence of standard quantum dynamics. This consequence can be argued for informally as follows. The Schrödinger equation effectively rotates a state-vector in Hilbert space continuously. The eigenstates of the m-operator will all be orthogonal to each other in this space. It follows that Schrödinger evolution from one m-eigenstate to the next must go through m-superpositions. But if collapse happens as soon as the first m-superposition arises then since the first m-superposition attributes effectively all amplitude to the initial m-eigenstate, the system will remain in that state forever. This consequence of the theory is refuted by our changing experiences.

[3] We might think of m-properties as measurement properties (at least upon precisifying the notion of measurement), or as meter properties (akin to meter readings), or as macroscopic properties (on a certain way of precisifying this notion), or as mental properties (on the view that gives a special role to the mind). More evocatively, we might think of them as Midas properties (everything m-properties touch turn to definiteness), or as (reverse) Medusa properties (everything m-properties see turn to stone).

[4] As before, such a constraint can be weakened. For example, we will later explore a hypothesis on which the m-property is a quantity, and the extent to which the property cannot superpose is a function of the amount of this quantity instantiated by a system.

This consequence can also be proven formally. Let the initial state of a (*any*) quantum system be $|\psi_0\rangle$ and the later state after time t be $|\psi_t\rangle$. The dynamical evolution of the system is described by a unitary operator U(t) that is a complex function of the system's Hamiltonian H such that $U(t) = e^{-iHt}$. We then have:

$$|\psi_t\rangle = U(t)|\psi_0\rangle$$

The "survival" probability P_s that the system will still be in the initial state at t is given by:

$$P_s = |\langle\psi_0|\psi(t)\rangle|^2 = |\langle\psi_0|e^{-iHt}|\psi_0\rangle|^2$$

From this we can derive:

$$P_s = 1 - (\Delta H)^2 t^2$$

Where:

$$(\Delta H)^2 = \langle\psi_0|H^2|\psi_0\rangle^2 - (\langle\psi_0|H|\psi_0\rangle)^2$$

This is an approximation valid for small t (obviously, because the expression becomes negative for $t > \Delta H$). You can obtain it by expanding the exponential as:

$$e^{-iHt} = 1 - iHt - \frac{1}{2}H^2t^2 + \cdots$$

and only keeping the terms up to order t^2 in the expression you get for the survival probability.

We may now define the Zeno time $Z = \frac{1}{\Delta H}$ so that:

$$P_s = 1 - \frac{t^2}{Z^2}$$

Now consider measurements separated by time intervals that become arbitrarily small. Instead of doing a single measurement after a time t, one performs N successive measurements at time intervals $\delta t = \frac{t}{N}$. In that case the expression for the survival probability becomes:

$$P_s^N = \left(1 - \frac{t^2}{N^2Z^2}\right)^N$$

It should now be clear what happens as t tends to zero and N tends to infinity: P_s tends to 1. So if the Schrödinger equation applies to m-properties, then the survival probability for a system that instantiates an m-property is 1. The strict stipulation therefore cannot work.

To weaken the constraints imposed on our m-property we can borrow a very recent idea from Kremnizer and Ranchin (2015) who themselves borrow ideas from GRW's spontaneous collapse model. To begin with, let's consider the GRW model.[5]

At random times, the wave function experiences a sudden jump, or "collapse", of the form:

$$\psi_t(x_1, x_2, \dots x_N) \rightarrow \frac{L_n(x)\psi_t(x_1, x_2, \dots x_N)}{\|L_n(x)\psi_t(x_1, x_2, \dots x_N)\|} \qquad (1)$$

Where $\psi_t(x_1, x_2, \dots x_N)$ is the wave function of the whole system at time t, immediately prior to the collapse. $L_n(x)$ is a linear operator equal to:

$$L_n(x) = \frac{1}{(\pi r_C^2)^{3/4}} e^{-(q_n - x)^2/2r_C^2} \qquad (2)$$

Where r_C sets the width of the localization process, and q_n is the position operator associated with the n-th particle of the system; the random variable x corresponds to the place where the jump occurs.

As well as expressing these equations in terms of the wave function, we can also express them in terms of the density matrix. The 1-particle master equation is:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H, \rho(t)] - T[\rho(t)] \qquad (3)$$

Where H is the Hamiltonian of the particle, and T[] represents the effect of the spontaneous collapses on the particle's wave function. In the position representation, this operator becomes:

$$\langle x|T[\rho(t)]|y\rangle = \lambda_{GRW}\left[1 - e^{-(x-y)^2/4r_C^2}\right]\langle x|\rho(t)|y\rangle \qquad (4)$$

The spontaneous collapse suppresses the off-diagonal elements of the density matrix, with a rate proportional to λ_{GRW} (equal to 10^{-16} probability per second for collapse), and to the distance between the off-diagonal elements: distant superpositions are suppressed faster than closer ones.

The many particle master equation is the generalization where an operator $T_i[]$, i = 1,2,...N appears for each particle. For ordinary matter one can separate the centre-of-mass motion from the internal motion. The reduced density matrix for the internal motion obeys the standard Schrödinger equation, while that for the centre of mass is equivalent to (3) where now the collapse rate is $N\lambda_{GRW}$, with N the total number of particles. This entails that the wave function collapses with a rate proportional to the *size* of the system.

[5] The presentation here summarises parts of a more detailed presentation found in XX.

In the GRW model, particles collapse *spontaneously* in the strong sense that no property can be said to cause the collapse. Kremnizer and Ranchin introduce a property as the cause of collapse by replacing GRW's rather arbitrary poisson process λ_{GRW} with a function the system's *integrated information*. We will say more about the concept of integrated information later [note: at the moment we don't].

For now, we speak in more general terms: λ_{GRW} can be replaced by a function of the system's *m-property* (whatever that m-property ultimately turns out to be). We may therefore replace equations (3) and (4) with:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H,\rho(t)] - f\left[M\left(\rho(t)\right)\right]\left(1 - e^{-(x-y)^2/\boldsymbol{M}[\varphi(\rho(t))]}\right)\langle x|\rho(t)|y\rangle \quad (5)$$

Here we have replaced phi with M (for m-property). There are two key differences between (5) and the GRW master equation. The first is that λ_{GRW} is replaced by $f\left[M\left(\rho(t)\right)\right]$ such that the rate of collapse is a function of the system's m-property. The second is the denominator. Where GRW define a Gaussian collapse function with a fixed width, we allow that the form of the collapse function may also be some function of the m-property. One motivation for doing this is as follows: Let the m-property be some physical quantity. If a system exhibits a small amount of the m-quantity then the system will undergo position collapse. If the collapse makes a significant effect on the system - significantly localising what was not already well localised - then this is likely to drastically change the system and potentially destroy it. This is an aspect of the project that will receive more attention in future research.[6]

While the KR model allows that collapse is a function of a particular property of the system, it does not capture the idea of an m-property as defined here, and is therefore not an instance of a triggered collapse theory. As we define them, triggered collapse theories postulate m-properties, which trigger collapse as a causal response to *entanglement*. In particular, by their nature m-property superpositions are unstable states so if a system tries to entangle an m-property (thereby superposing the m-property), then the m-property causally responds by collapsing the wavefunction. The question, then, is whether we can modify the KR dynamics to capture this stipulation.

Our proposed master equation takes equation (6) but replaces $f\left[M\left(\rho(t)\right)\right]$ with a different function $j\left[M\left(\rho(t)\right)\right]$. While f is a measure of the m-quantity of a system j is a measure of m-quantity distributions across superposition components of the system. In particular it is a measure

[6] As a suggestion for replacing the GRW Gaussian: let the collapse function be Gaussian but let the width of the Gaussian be a function of the m-property. In particular, if a system's m-quantity is small then the system's particles localise only very slightly. As the system grows the localisation becomes stronger, and the particles are eased into near-eigenstates of position. A further issue concerns superpositions of low m-values. The superposition components will compete to be collapse centres. This competition will have to be resolved.

of the distinctness of superposition components. Mathematically, this measure can be modeled by a non-Hermitan operator on the Hilbert space. In particular, we can think of it as a linear map from superpositions of m-eigenvectors to m-eigenvectors. The operator is defined so that superpositions of *more distinct* eigenstates are mapped to larger eigenvalues. The frequency of collapse is then a function of these eigenvalues.

We also intend to make a similar stipulation about function g (in the denominator). For we want the m-value variance among superposition components to affect not only the frequency of collapse but also the *severity* of collapse. So if the variance is low collapse has little effect, whereas if variance is high then collapse localises the particles in the system to a greater extent. Given these stipulations, as the m-quantity grows and varies in the various initial components, the frequency and severity of collapse is raised. In initial evolution of the m-quantity, superpositions of low m-values are steered towards definite high m-values. In measurement scenarios, superpositions of m-values with high variation are immediately achieved but are also highly unstable and collapse immediately with a severity that puts the m-system back into a definite state. Our master equation is therefore:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H,\rho(t)] - j\left[M\left(\rho(t)\right)\right]\left(1 - e^{-(x-y)^2/\boldsymbol{g}\left[M\left(\rho(t)\right)\right]}\right)\langle x|\rho(t)|y\rangle \quad (6)$$

While the mathematical details of j (and g) require more work, we believe that (6) offers a useful starting point for further development. For example, this equation can be used to specify the dynamics of the wavefunction on the triggered-collapse interpretation associated with any given m-property. Note that in this dynamics there is no mention of measurement at all (that step-ladder has been kicked away).

We can illustrate this dynamics by choosing a particularly simple m-property. We can suppose that *m-particles* are a rare sort of fundamental particle whose position is constrained to be always determinate. This yields a triggered-collapse interpretation for the m-property defined as follows: particles whose position wave-functions are Gaussians, such that amplitude is bunched in a small region, will be stable: probability for collapse will be extremely low. That's because variance among superposition components is extremely low: almost all amplitude is distributed among position eigenstates that correspond to points in a small region. On the other hand, if the particle is in a superposition of multiple Gaussians, where the Gaussians are far apart, the state is unstable, and probability for collapse is high.

On this interpretation, collapses will occur when superposed systems interact with m-particles. Suppose we put an m-particle in position p inside one slit of a double-slit experiment, in such a way that (superpositions aside) one would expect it to move to position p' if and only if the electron passes through that slit. Without the m-particle, the position of the electron at the timing of passing through the locations of the slits would be a superposition

$\frac{1}{\sqrt{2}}$(|p> + |q>), where q is the location of the second slit. If a particle other than an m-particle were inside the slit, then when the particle enters the slit we would expect its position to become entangled with that of the electron, yielding a superposed state $\frac{1}{\sqrt{2}}$(|p>|p'> + |q>|q'>). However, if an m-particle is present inside the slit, the system will collapse into either |p>|p'> or |q>|q'> with probability 0.5 for each. In effect, the m-particle is serving as a measuring device.

Though perhaps fanciful, this is a perfectly coherent dynamics to postulate. Furthermore, under certain assumptions, the hypothesis is not too far away from predicting the results to date of quantum measurement. We need only suppose that m-particles are just rare enough (say, one in every 10^8 particles is an m-particle) that (i) no quantum-mechanical experiments have been performed on isolated systems containing m-particles and (ii) m-particles are always present in our measuring devices (or in human perceptual processes), then this hypothesis might have a chance of reproducing the results to date of quantum measurements.

We are not offering the m-particle hypothesis as a serious hypothesis. There is not much positive reason to believe in a special sort of particle satisfying (i) and (ii). If there were one, we could quickly attempt to falsify the hypothesis by performing a double-slit experiment in which the particle is fired through the slits. If typical interference effects result, this would indicate a superposed position for the particle and rule out the hypothesis above.

Still, the m-particle hypothesis illustrates the way that triggered-collapse interpretations work. And the model generalizes to a wide variety of other potential m-properties. In principle there is a triggered-collapse interpretation of quantum mechanics for any m-property at all. However, a wide variety of these interpretations will be ruled out by existing empirical data. As in the case above, two minimal 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 self-identical: on these interpretations, measurement devices and human perceptual processes would never collapse, and neither measurement nor perception would yield determinate results.

[[[Perhaps one could argue that we are simply experiencing one component of a superposed perceptual state; but this way of thinking leads naturally to a many-worlds interpretation rather than a collapse interpretation. Or perhaps one could argue that my data require only that *one's own* perceptual processes yield definite results, or even that one's *current* perceptual processes yield definite results (one can treat other humans and even past versions of myself in the way one would treat superposed external measuring devices); but we will stay with the standard non-solipsist assumption that assumes that there is nothing special about oneself and the current time.]]]

For an interpretation 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: [need some better ones, suggestions welcome!]

(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) 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).

(4) 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.

Questions and Objections

(1) Are triggered-collapse models empirically testable?

A nice feature of triggered-collapse models is that different models involving different m-properties make different predictions. This is a consequence of the fact that in principle it is possible to test whether a system is in a superposition of an m-property, for any m-property whatsoever. To do this one can use an interferometer, which detects interference between superposed quantities in much the way that a double-slit experiment detects interference between superposed positions. In practice it is extraordinarily difficult to set up these measurements for complex m-properties (for reasons, see Albert 1992), but the measurements are possible in principle. Using these methods one could falsify any triggered-collapse interpretation by detecting superpositions of the associated m-property. By performing enough of these tests, one could determine precisely which m-properties are never superposed, and one could use these results to infer which are the underlying m-properties that serve as triggers.

So in principle (though not yet in practice), we can determine empirically whether a triggered-collapse interpretation is correct, and if so which such interpretation is correct. If it turns out that every physical property can be superposed, then no triggered-collapse interpretation is correct. If it turns out that some physical properties can never be superposed, however, then we will have strong evidence in favor of a triggered-collapse interpretation. Sufficient testing will then give us strong guidance as to precisely which such interpretation is correct.

To date, the relevant empirical data is limited. Quantum effects suggest that m-properties cannot be fundamental physical properties. Superposition effects have been demonstrated for positions in molecules as large as buckminsterfullerene (C_60 and C_70), so it seems likely that nontrivial m-properties will attach to systems more complex than this. As superposition effects are demonstrated more widely, triggered collapse theories will become more constrained. One can hope that eventually it will be possible to test for superpositions involving human brains. If

experiments rule out superpositions here (or at some lower level), then we will have strong evidence for a triggered-collapse theory. If experiments demonstrate superpositions at this level, however, then there will be good reason to reject such theories, and perhaps to reject any collapse theories in favor of no-collapse theories such as many-worlds and hidden-variable interpretations.

(2) 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.

[**I suggest we delete this section. In our equation collapses are to position, no matter the m-property. And they are Gaussian collapses so the violations of energy conservation are too minimal to be of interest. The form of the Gaussians are now a function of the m-property, but I'm thinking we define this function so that GRW collapses are the most localised collapses possible. **]

(3) What about the tails problem?

On our model m-property-induced collapse involves multiplying the position wavefunction of the system by a Gaussian. This leaves behind tails: although the collapse centre acquires the majority of the mod-square amplitude, the system's superposition retains (low-amplitude) tails. This is a problem if those low amplitude tails exhibit structure resembling the structure in the collapse centre for then the model would be many-worlds in disguise (McQueen 2015).

Wallace (XX) and Vaidman (XX) have shown that the structures in the tails are (eventually) significantly distorted by the collapses. McQueen (2015 XX) argues that if Wallace is right that such distortions are (at least to begin with) relatively minor, the tails problem remains. However, Vaidman suggests that observers in the tails almost immediately disintegrate. If they disintegrate quickly enough, say, before self-awareness comes into play, then arguably the tails problem is solved. However, we simply do not currently have the complex calculations that would enable us to resolve this issue.

If anything like Wallace's estimates are accurate, then we require a solution to the tails problem. Chalmers (2012) defends GRW against the tails problem by appeal to what he calls *spatial functionalism*. McQueen (2015 XX) has responded that this approach presupposes controversial principles about when physical systems exhibit consciousness. This is a problem for GRW since one of the primary aims of GRW is to solve the measurement problem without having to appeal (at any point) to conscious experience. This is perhaps less of a concern on the present account. The issue requires more attention in future research.

(4) 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 amplitude of nonzero states creeps above zero, it will immediately collapse back to zero? Figure this out!

[**I suggest we delete this section since it is discussed (and arguably resolved) in the discussion above**]

(5) Can triggered-collapse models 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.

[**I wonder whether we should delve into such a complicated issue, or just reference the progress on relativising GRW by e.g. Tumulka and Bedingham? There's also the issue, raised by Albert (2015), that relativistic quantum field theory, in order to be narratable, needs a Lorentz interpretation of special relativity that does have a privileged frame after all.**]

(6) 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. See the following section.

[**Hasn't Schaffer written about non-fundamental properties in fundamental laws?]

Consciousness as an 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 discuss shortly) 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 (significantly) 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 $\frac{1}{\sqrt{2}}$(|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 $\frac{1}{\sqrt{2}}$ (|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 $\frac{1}{\sqrt{2}}$ (|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:

(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 formalism so that states of consciousness are included in the underlying space that yields the configuration space within which the wave function sits. [How exactly would this work?] 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. [Figure this out!]

[**Why is it not enough that phi is (straightforwardly) represented in the wave-function? If phi is physical and we assume substance dualism then consciousness can be treated like a field that is determined by phi. But it could also exist in its own pre-existing phenomenal space such that it couples with physical systems when those systems exhibit enough phi. **]

(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.

(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 consciousness, 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.

(4) What about property P?

An opponent might object that even on a consciousness-collapse view, there will need to be fundamental psychophysical laws connecting property P 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.

[Say something other consciousness-collapse views: Stapp, Hodgson, Hameroff and Penrose, Wigner?]