The Conway-Kochen Free Will Theorem

Tarun Menon

September 2009

1 Introduction

John Conway and Simon Kochen say that their recent Free Will Theorem (henceforth, FWT) is "the culmination of a series of theorems about quantum mechanics that began in the 1960s" (Conway & Kochen, 2009, p. 226), theorems which place constraints on the interpretation of quantum mechanics by proving that certain sets of physical assumptions are incompatible with the quantum formalism. The FWT is in fact a combination of two such theorems: Bell's Theorem (Bell, 1964) and the Kochen-Specker Paradox (Kochen & Specker, 1967). Conway and Kochen believe that the FWT has novel implications that go beyond the results of these two theorems. However, the extent of this novelty has been challenged in Bassi and Ghirardi (2007) and Tumulka (2007), both reactions to the first publication of the FWT in Conway and Kochen (2006). In response, Conway and Kochen have developed a strengthened version of the theorem (Conway & Kochen, 2007, 2009) that they claim illustrates the import of their theory in a more perspicuous manner. In this paper, I will examine whether this version of the FWT presents us with any genuinely new constraints on interpretation. In particular, does it rule out any among the interpretations of quantum mechanics currently considered plausible?

2 The Free Will Theorem

The theorem is proved from three axioms, pleasingly titled SPIN, TWIN and MIN. According to the SPIN axiom, the squared spin components of a spin 1 particle along three orthogonal directions can be simultaneously measured, and the measurement result will be 1 along two of the three directions and 0 along the third. That is, if x, y and z are the three directions, and $\hat{S}_x$, $\hat{S}_y$ and $\hat{S}_z$ are the spin operators along these directions, then the squared operators $\hat{S}_x^2$, $\hat{S}_y^2$ and $\hat{S}_z^2$ commute, and measuring the corresponding observables simultaneously will give us either $\{1, 1, 0\}$, $\{1, 0, 1\}$ or $\{0, 1, 1\}$ as the set of measurement results along x, y and z respectively. The second axiom, TWIN, states the possibility of an EPR-type experiment. Two spin 1 particles a and b can be prepared in a state such that if an observer A measures the squared spin components of a along the directions x, y and z, and a distant (space-like separated) observer B measures the squared spin of b along a direction w, then if w is parallel to either x, y or z, B's measurement will agree with A's measurement along that direction. In other words, the particles can be prepared in a twinned state so that measurements of squared spin along the same direction are perfectly correlated.

Both SPIN and TWIN are straightforward consequences of quantum mechanics. The third assumption, MIN, is potentially more controversial. It combines assumptions about the free will of the experimenters, special relativity and the impossibility of backward causation. Take any set of 33 directions in space that are related according to the Peres configuration (see p. 227 of (Conway & Kochen, 2009) for a description of this configuration of directions). The MIN axiom says that in the EPR experiment described above, the observer B can freely choose any of these 33 directions to measure the squared spin of particle b, and particle a's response is independent of this choice. Similarly, observer A can

freely choose between the 40 orthogonal triples formed by completing orthogonal pairs in the Peres configuration in order to measure the spin components along these directions of particle a, and particle b's response is independent of this choice. For a choice to be free just means that it is not functionally dependent on any information in its past half-space relative to any inertial frame. It is assumed that the responses of the particles are independent of the choices made at the opposite arm of the experiment because the two observers are space-like separated. If there was some influence from, say, A's choice of experimental setting to the response of particle b to B's experiment, we would have to countenance backward causation in some inertial frame.

Using these axioms, one can prove that the response of the spin 1 particle a to A's experiment must be free in the sense already discussed. That is, the response must be functionally independent of any information in its past relative to any inertial frame. This is the Free Will Theorem. Conway and Kochen offer this rough colloquial formulation of the FWT: "[...] if indeed we humans have free will, then elementary particles already have their own small share of this valuable commodity." (Conway & Kochen, 2009, p. 226) To be more precise, the response of particle a is not completely free but "semi-free", since it is constrained by the requirement of appropriate correlation with the response of particle b set by the TWIN axiom.

The authors prove the FWT by contradiction. They assume that the particle a's response to the measurement of its squared spin components along three orthogonal directions is indeed a function of past information in some arbitrary inertial frame. This is called the "functional hypothesis". They then prove that this hypothesis has the consequence that one can construct a function with the 33 directions of the Peres configuration as the domain and $\{0, 1\}$ as the range such that in every orthogonal triple of directions only one of the directions is

mapped to 0. This is just a version of the Kochen-Specker paradox using the simplified construction from Peres (1991) rather than Kochen and Specker's original proof. Anyone familiar with that result should know that there is no function satisfying the specified conditions. So the axioms SPIN, TWIN and MIN are incompatible with the functional hypothesis. The response of particle a must be free.

It is worth noting that Conway and Kochen want their theorem to be as free of theoretical assumptions as possible. They advise that seemingly theoretical terms such as "spin 1 particle" be operationally construed as referring to facts about the behavior of macroscopic measuring devices, such as spots on a screen. This helps further specify the assumption about relativistic locality embedded in the axiom MIN. The independence of particle a's response to the choice of direction made by observer B amounts to the claim that a macroscopic object (in this case, A's measuring device) cannot have its state changed by altering a spatially separated macroscopic object (in this case, B's experimental apparatus). In Michael Redhead's careful analysis of different locality assumptions in quantum mechanics (Redhead, 1987), this is the one labeled LOC_4 (ibid., p. 91).

Conway and Kochen believe that their theorem places several constraints on the interpretation of quantum mechanics if it is to be reconciled with relativity. They say that the FWT shows there can be no relativistic deterministic theory of nature, ruling out the possibility of a relativistically kosher version of Bohm's hidden variable interpretation (Bohm, 1952). In addition, and more surprisingly, they say that the theory also rules out the possibility of any stochastic theory that provides a mechanism for the reduction of the wave function such as the GRW theory (Ghirardi, Rimini, & Weber, 1986). This latter claim might seem especially surprising (maybe even straightforwardly false) given the recent

construction of a relativistic model of GRW theory using the "flash" ontology (Tumulka, 2006). However, this model does not yet fully incorporate interactions, and Conway and Kochen take their theorem to entail that it never will.

3 Implications for Hidden Variable Theories

Does the FWT stake out new ground in ruling out hidden variable theories? The authors point out that the strategy of "contextuality" used by hidden variable interpretations to defeat past no-go theorems (such as the Kochen-Specker theorem) will not work against the FWT. According to contextual theories, the response of particle a to the squared spin experiment along any particle direction depends on the entire orthogonal triple of directions chosen by the experimenter. The choice of orthogonal triple (x, y, z) could influence the particle's response along direction z either by altering the value of the squared spin component along that direction, or by determining which one of several different observables associated with the squared spin operator is measured by the experiment. Redhead (1987) calls the former approach "environmental contextuality" and the latter "ontological contextuality". Both strategies are ruled out by the FWT. If some sort of interaction with the experimental apparatus fixes either the value of the particle's squared spin, or fixes which among several pre-existing values will be displayed, then the particle's response is dependent on this interaction and hence not free, in contravention of the FWT.

This incompatibility with even contextual hidden variable theories, such as Bohm's theory, might seem to be a novel result. However, it is important to note that the axioms of the FWT contain a locality assumption. It has been known since Bell proved his famous theorem that *local* hidden variable theorems (whether contextual or not) are incompatible with quantum mechanical assumptions. As I mentioned above, the locality assumption embedded in MIN

is Redhead's LOC_4. Redhead argues that this locality assumption entails the Bell inequality if we also assume determinism (ibid., pp. 90-96). And, of course, the Bell inequality is incompatible with the quantum formalism. So the prohibition on local deterministic theories is nothing new.

A theorem may be novel not just in virtue of its conclusion, but in virtue of its axioms. In particular, a theorem that proves a previously known conclusion using weaker assumptions might well be of considerable value. Redhead's argument relies on demonstrating that quantum mechanics violates the Bell inequality. Is this violation derivable from SPIN, TWIN and MIN, or do we need to make additional substantive assumptions? Assuredly not. The Bell inequality is a statistical claim, but no statistical assumptions have gone into the proof of the FWT. The "independence" spoken of in the MIN axiom is *functional* independence, not *probabilistic* independence. I stress this point because a conflation of these notions seems to underlie some of Conway and Kochen's argumentation, and is also evident in the responses of some of their critics. I will have more to say on this later. For now, suffice it to say that a derivation of the Bell inequality requires stronger assumptions than those that go into the FWT, so Redhead's result does not challenge the novelty of the theorem.

However, there is another result in Redhead's book that does. The Kochen-Specker theorem, unlike Bell's theorem, does not involve statistical assumptions. As we have seen, it does not rule out contextual hidden variable theories, but Redhead argues that in the face of the Kochen-Specker theorem any contextual hidden variable theory must violate one (or both) of two locality conditions, which he calls ontological locality (OLOC) and environmental locality (ELOC) (ibid., pp. 139-142). OLOC says that the observable of a subsystem measured by an experiment is independent of the observable measured for a distant subsystem that is part of the same composite system, and ELOC says that the values

of the observables of a subsystem cannot depend on the setting of a distant apparatus that is part of the measurement context for the composite system. What matters here is that a violation of either condition entails a violation of the sort of locality condition assumed by Conway and Kochen.

If OLOC is violated in the TWIN experiment, the observable measured on one arm of the experiment depends on the measurement context of the composite system, including the observable measured on the other arm. Altering the measurement setting on B's arm changes the composite observable being measured. This alters the observable being measured on A's arm, and therefore particle a's response. If ELOC is violated, the value of the observable (but not the observable itself) measured by A depends on the measurement setting chosen by B. In both cases, the macroscopic spots on a screen on A's arm will depend on the B's measurement setting, a violation of LOC_4. So LOC_4, and therefore MIN, entails both ELOC and OLOC. The converse does not hold: the conjunction of ELOC and OLOC does not entail MIN, since neither assumption says anything about the choice of experiment being free. So Redhead has proved the impossibility of deterministic hidden variable theories (contextual or not) using strictly weaker assumptions than Conway and Kochen. Both Redhead's result and the FWT rely on the Kochen-Specker theorem, but Redhead's locality assumptions are weaker. The FWT apparently does not tell us anything new about the impossibility of deterministic hidden variable interpretations.

In fact, a more careful look at the axioms used to prove the FWT shows that the theorem does not have much to say about such interpretations at all. The axiom MIN contains the free will assumption that the choices of experimental setting made by the two experimenters is functionally independent of all past information. Any deterministic theory that eschews mind-body dualism will reject this assumption. If the experimenters' decisions ultimately come down

to interactions between particles in their brains, then these decisions would be under the purview of the deterministic theory and so perforce be functions of past information. The FWT, then, begs the question against the Bohmian (or any other proponent of a deterministic hidden variable theory). So unlike Redhead's theorem, the FWT places no substantive constraints on the possibility of deterministic theories (unless ruling them out by *fiat* is considered substantive).

Conway and Kochen do offer some arguments to justify the free will assumption. They say it would be hard to take the results of science seriously if the choices of experimenters were determined. In particular: "Physical induction, the primary tool of science, disappears if we are denied access to random samples." (Conway & Kochen, 2006, p. 1466) Their critics make similar arguments. Here's Tumulka: "It seems to me that any theory violating the freedom assumption invokes a conspiracy and should therefore be regarded as unsatisfactory." (Tumulka, 2007, p. 188) Bassi and Ghirardi say "no-one is willing to deny [the free will assumption]" (Bassi & Ghirardi, 2007, p. 173) In fact, none of the published criticisms of the FWT that I have seen have challenged this assumption. What is behind this almost uniform avowal?

I think the conflation mentioned above, between functional and probabilistic independence, is to blame. The quote from Tumulka, branding the rejection of freedom conspiratorial, is instructive. Here Tumulka seems to be rejecting not determinism, the view that the complete description of field values on a space-like hypersurface along with the laws fixes the fields throughout the entire space-time manifold, but what has been called *superdeterminism*, the view that the initial conditions of the universe were specially set up to produce certain correlations that are not robust under changes of those initial conditions. In a superdeterministic account of the TWIN experiment, there is a non-nomic correlation between the physical state of the two-particle system prior to mea-

surement and the chosen detector settings, so that we get the statistics predicted by quantum mechanics even though nature is potentially local and deterministic. This is not an automatic consequence of determinism. It depends on the particular probability distribution placed over the state space of your deterministic theory. If the distribution has the property (for all states λ and detector settings d_A and d_B) that $P(\lambda|d_A, d_B) = P(\lambda)$, then the world, though deterministic, is not superdeterministic or conspiratorial. Bohmian mechanics is one example of a deterministic theory that is not superdeterministic. Even though the deterministic laws entail that the choice of detector settings in any particular instance of the experiment is functionally dependent on the past, according to the distribution postulate the settings are probabilistically independent of the hidden state of the system being measured. A similar argument applies against Conway and Kochen's claim that a deterministic theory denies us access to random samples. Determinism is a feature of the non-probabilistic laws while randomness is a feature of the probability distribution placed over the deterministic state space. For a sample to be random is just for there to be no correlations (under this distribution) between the sample selection mechanism and the population from which the sample is selected.

4 Implications for Stochastic Theories

We have seen that the FWT does not present us with a new verdict on the possibility (or otherwise) of deterministic theories. But what about the authors' claim that the theory also rules out stochastic theories (such as the GRW theory) that provide a mechanism for reduction? If this claim is true, then it is certainly a novel result. As mentioned, relativistic models of GRW theory are being developed, with Tumulka's relativistic GRW model with flashes (rGRWf) being a prominent example. If the FWT does indeed have the touted consequence,

this budding research program will die before it blooms.

The strategy for ruling out stochastic theories is essentially to use a trick to convert them into something like deterministic theories for which the functional hypothesis is true. Conway and Kochen's initial argument (presented in Conway and Kochen (2006)) was that any random element in the mechanism for reduction could be specified beforehand, thus making it information about the past. This is analogous to making all the die rolls in a monopoly game before the game actually starts, and then using that previously fixed information to play the game. But if this information is relegated to the past, the FWT proves that it is not relevant to the particle's response. So the stochastic mechanism cannot be responsible for the particle's "decision". As an example of this procedure, take the random elements in the GRW model – the position of the "hits" and the time between them – and just provide all the information about the distribution of these elements prior to the experimenters' choices. Then, by the FWT, this information does not determine the particles' behavior. A mere change in the time at which the stochastic information is "inserted" makes no fundamental difference to the theory, so the random elements in the GRW model are not sufficient to fix the particles' responses.

Tumulka (2007) responds that the strategy of pre-generating random information outlined by Conway and Kochen will not work on rGRWf because the distribution of flashes determining the response of a particle in an EPR-type experiment depends on the external fields generated by the choices of directions made by experimenters on both arms. If this distribution were given in advance, then in order to preserve this dependence, the external fields must also be given in advance (unless we allow the future to influence the past, which Conway and Kochen rule out). But these fields cannot be given in advance, because by hypothesis the experimenters' choices are free of past information, and the fields

are determined by those choices.

Conway and Kochen (2009) modify the earlier argument to get around Tumulka's response. The MIN assumption supposes that there are 1320 (40×33) different combinations of directions that the two experimenters in the TWIN experiment might choose. Instead of pre-generating information about flashes that depends on the particular choice they do make, pre-generate flash distributions for every one of the 1320 possibilities. Then, when the choice of directions has actually been made, use that information in conjunction with the 1320 pre-generated flash distributions to determine the actual response of the particle. But that would be a violation of the FWT, since the particle's response would not be independent of past information, so it is impossible. This response is problematic, and its problems highlight the real incompatibility between rGRWf and the FWT, viz. that Tumulka's theory rejects MIN. Once information about the detector settings is available, one out of the 1320 pre-generated flash distributions is picked to determine particle a's behavior. But to narrow it down to *one* out of the 1320 possibilities, we need information about the setting chosen at both arms of the experiment. This means the response of particle a will depend on the setting chosen by observer B, in violation of the MIN axiom.

As (Tumulka, 2007) already pointed out, it is not (as Conway and Kochen suppose) that the FWT rules out some future extension of rGRWf to include interactions. The model, as it is now, rejects MIN, and so is incompatible with the FWT. According to the model, the flash distribution that determines (or, more accurately, instantiates) particle a's response will depend on the choices made by both experimenters. So much the worse for rGRWf, one might think. As Conway and Kochen argue, the rejection of the locality assumption in MIN implies that one is willing to countenance backward causation in some inertial frame, surely not a congenial result. But all this talk of "cause" and "influence"

can be misleading with a theory like Tumulka's. The genius of the theory is paring down the ontology to only include flashes. One could talk of flashes "influencing" one another, but this should be recognized as a mere frame-dependent *façon de parler*. As Tim Maudlin points out (Maudlin, 2008), strictly speaking Tumulka's theorem only delivers the conditional probabilities of certain flashes given other flashes. There is no temporal directionality to conditional probabilities, so we don't get paradoxes by switching to a different inertial frame, i.e. "the notion of temporal or causal priority of one measurement over the other never appears in this theory." Once we have abandoned a robust notion of causal influence between the fundamental elements of reality, the locality assumption in MIN is no longer plausible. Norton (2007) has more on why this notion of causation has no place in fundamental physics.

Conway and Kochen are uncomfortable with this kind of talk. They say, "MIN does not deal with flashes or other occult events, but only with the particles' responses as indicated by macroscopic spots on a screen, and [the causal order of] these [is] surely not frame-dependent." (Conway & Kochen, 2009, p. 232) But this just begs the question against rGRWf, according to which the flashes are not "occult" at all, they are all that really exists. Macroscopic spots on a screen are just large collections of flashes, so if the flashes have no frame-independent causal order or absolute relation of influence between them, neither do the spots. The postulation of non-causal macroscopic correlations by rGRWf should not be a difficulty for Conway and Kochen, since they do it themselves. Their own approach to the correlation observed in the TWIN experiments is to offer no explanation at all, just accept it as brute fact (Conway & Kochen, 2009, p. 230). So they have already rejected the assumption that correlations must be explained by some metaphysically robust sense of causation. Why must the alternative be no explanation at all? Why not allow a theory like rGRWf, which

explains the theory using conditional probabilistic constraints derived from its dynamics? Surely this is more satisfactory than just asserting the constraints as primitive.

5 Conclusion

Have we learned anything new from the FWT? I think the answer is a yes, but we have not learned what was advertised. Conway and Kochen say, "Granted our three axioms, the FWT shows that nature itself is non-deterministic." (Conway & Kochen, 2009, p. 231) This is true, but trivially so, since non-determinism is presumed by the MIN axiom. They go on to say, "Moreover, the FWT has the stronger implication that there can be no relativistic theory that provides a mechanism for reduction." (ibid., p. 231) This is not true, since MIN is too strong a condition to capture the requirement of relativistic invariance. In addition to Lorentz-invariance, MIN presumes that a theory that attempts to explain correlations can only do so using a metaphysically robust, frame-independent notion of causality. Nothing in relativity theory requires this. What the FWT makes apparent more than other no-go theorems is that any relativistic theory that aspires to explain non-local correlations must abandon this notion of causality. Appropriately construed, rGRWf does just this. There is matter of fact about which flash "causes" which. There are merely symmetric relationships of probabilistic dependence between flashes.

The FWT derives a contradiction from the following assumptions: (i) freedom, (ii) TWIN correlations, (iii) Lorentz covariance, (iv) the functional hypothesis (i.e. the particle's response is a function of past information) and (v) robust causation. Conway and Kochen reject assumption (iv), decreeing that particle's have a modicum of free will. But this forces them into accepting "semi-freedom" – the correlation between particles a and b –as a brute fact.

We have seen that there is a more attractive resolution available. If we reject assumption (v) we can construct relativistic theories that actually explain non-local correlations. This is the true lesson of the theorem.

References

Bassi, A., & Ghirardi, G. (2007). The Conway-Kochen Argument and Relativistic GRW Models. *Foundations of Physics*, *37*, 169-185.

Bell, J. S. (1964). On the Einstein Podolsky Rosen Paradox. *Physics*, *1*, 195-200.

Bohm, D. (1952). A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables. *Physical Review*, *85*, 166-179.

Conway, J., & Kochen, S. (2006). The Free Will Theorem. *Foundations of Physics*, *36*, 1441-1473.

Conway, J., & Kochen, S. (2007). Reply to Comments of Bassi, Ghirardi, and Tumulka on the Free Will Theorem. *Foundations of Physics*, *37*, 1643-1647.

Conway, J., & Kochen, S. (2009). The Strong Free Will Theorem. *Notices of the American Mathematical Society*, *56*, 226-232.

Ghirardi, G., Rimini, A., & Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Physical Review D*, *34*, 470-491.

Kochen, S., & Specker, E. P. (1967). The Problem of Hidden Variables in Quantum Mechanics. *Journal of Mathematics and Mechanics*, *17*, 59-87.

Maudlin, T. (2008). Non-local correlation in quantum theory: How the trick might be done. In W. L. Craig & Q. Smith (Eds.), *Einstein, Relativity and Absolute Simultaneity* (p. 156-179). Routledge.

Norton, J. (2007). Causation as Folk Science. In H. Price & R. Corry (Eds.),

Causation, Physics, and the Constitution of Reality (p. 11-44). Clarendon Press.

Peres, A. (1991). Two simple proofs of the Kochen-Specker theorem. *Journal of Physics A: Mathematical and General*, *24*, L175-L178.

Redhead, M. (1987). *Incompleteness, Nonlocality, and Realism.* Clarendon Press.

Tumulka, R. (2006). A Relativistic Version of the Ghirardi-Rimini-Weber Model. *Journal of Statistical Physics*, *125*, 821-840.

Tumulka, R. (2007). Comment on "The Free Will Theorem". *Foundations of Physics*, *37*, 186-197.