Studies in History and Philosophy of Modern Physics 57 (2017) 98–103

Contents lists available at ScienceDirect

Studies in History and Philosophy of Modern Physics

journal homepage: www.elsevier.com/locate/shpsb

On the notion of free will in the Free Will Theorem ☆

Klaas Landsman

Institute for Mathematics, Astrophysics, and Particle Physics, Faculty of Science, Radboud University, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 17 October 2015
Accepted 2 November 2016
Available online 24 November 2016

Keywords:
Free Will Theorem
Local miracle compatibilism

ABSTRACT

The (Strong) Free Will Theorem (FWT) of Conway and Kochen (2009) on the one hand follows from uncontroversial parts of modern physics and elementary mathematical and logical reasoning, but on the other hand seems predicated on an undefined notion of free will (allowing physicists to "freely choose" the settings of their experiments). This makes the theorem philosophically vulnerable, especially if it is construed as a proof of indeterminism or even of libertarian free will (as Conway & Kochen suggest).

However, Cator and Landsman (*Foundations of Physics* 44, 781–791, 2014) previously gave a reformulation of the FWT that does not presuppose indeterminism, but rather assumes a mathematically specific form of such "free choices" even in a deterministic world (based on a non-probabilistic independence assumption). In the present paper, which is a philosophical sequel to the one just mentioned, I argue that the concept of free will used in the latter version of the FWT is essentially the one proposed by Lewis (1981), also known as 'local miracle compatibilism' (of which I give a mathematical interpretation that might be of some independent interest also beyond its application to the FWT). As such, the (reformulated) FWT in my view challenges compatibilist free will à la Lewis (albeit in a contrived way via bipartite EPR-type experiments), falling short of supporting libertarian free will.

1. The Free Will Theorem

The Free Will Theorem (FWT) of Conway and Kochen (2006, 2009) shows that some small and uncontroversial corner of quantum mechanics (i.e., the response of massive particles with spin one to measurements of their spin) combined with a rather weak locality condition suggested by Einstein's theory of special relativity (which effectively forbids superluminal signaling), is incompatible with the conjunction of determinism and the ability of experimentalists to "freely choose" the directions along which they measure spin. The FWT was published in two versions, of which the second, called the *Strong* Free Will Theorem by Conway and Kochen, has superseded the first (which may therefore be discarded). Conway and Kochen (2009, p. 226) paraphrase their theorem in the following way:

> 'if indeed we humans have free will, then elementary particles already have their own small share of this valuable commodity. More precisely, if the experimenter can freely choose the directions in which to orient his apparatus in a certain measurement, then the particle's response (to be pedantic—the universe's response near the particle) is not determined by the entire previous history of the universe. (…) our theorem asserts that if experimenters have a certain freedom, then particles have exactly the same kind of freedom. Indeed, it is natural to suppose that this latter freedom is the ultimate explanation of our own. (…) Granted our three axioms [i.e., the physical ones and freedom of choice], the Free Will Theorem shows that nature itself is nondeterministic.'

It is clear from Conway's recent biography (Roberts, 2015) that the authors saw their FWT as a major contribution to science (perhaps even to philosophy), and indeed it has generated considerable publicity. Part of this interest has been rather critical (cf. Bassi & Ghirardi, 2007; Cator & Landsman, 2014; Goldstein, Tausk, Tumulka, & Zanghi, 2010; Hemmick & Shakur, 2012; Hermens, 2014, 2016; 't Hooft, 2007; Wüthrich, 2011), mainly on the following grounds:

1. Lack of novelty compared with the famous paper by Bell (1964), whose assumptions and conclusions are at least quite similar to

☆Dedicated to Henk Barendregt, on the occasion of his official retirement (October 1, 2015).

E-mail address: landsman@math.ru.nl

http://dx.doi.org/10.1016/j.shpsb.2016.11.001
1355-2198/

those of the FWT, although the underlying proofs are mathematically distinct from those in the FWT.
2. Lack of novelty even within its own terms: almost identical results, even based on very similar mathematical reasoning, had previously been published by Heywood and Redhead (1983), Stairs (1983), Brown and Svetlichny (1990), and Clifton (1993).
3. Circularity, in that indeterminism is presupposed (namely in the assumption that 'experimenters have a certain freedom') instead of derived.

I only discuss these criticisms here in so far as they justify my own contribution; my take is that all of the above criticism is deserved, but that nonetheless the FWT is an interesting result, which triggers further discussion (of which the present paper is an instance).

1. The difference between earlier literature (of which, incidentally, Conway and Kochen only cite Heywood and Redhead) and the FWT is almost exclusively one of emphasis, namely on free will. Given this emphasis, it is striking that one looks in vain for serious philosophical analysis in Conway and Kochen (2006, 2009). All one finds is:

> 'The tension between human free will and physical determinism has a long history. Long ago, Lucretius made his otherwise deterministic particles swerve unpredictably to allow for free will. It was largely the great success of deterministic classical physics that led to the adoption of determinism by so many philosophers and scientists, particularly those in fields remote from current physics. (This remark also applies to "compatibilism", a now unnecessary attempt to allow for human free will in a deterministic world.)' (Conway & Kochen, 2009, p. 230).

Also elsewhere, one finds little respect for the philosophical debate on free will, e.g.,

> 'Compatibilism is an old viewpoint from previous centuries when philosophers were talking about free will. The were accustomed to physical theory being deterministic. And then there's the question: How can we have free will in this deterministic universe? Well, they sat and thought for ages and ages and ages and read books on philosophy and God knows what and they came up with compatibilism, which was a tremendous wrenching effect to reconcile 2 things which seemed incompatible. And they said they were compatible after all. (…) But in my view is it's really nonsense. (Conway, quoted in Roberts, 2015, p. 361).

Thus the main goal of the present paper is to relate the FWT (whatever its novelty compared to its predecessors) to the philosophy of free will. However, the (negative) relationship we are going to establish will be with compatibilist free will à la Lewis, as opposed to the (positive) relationship with libertarian free will envisaged (but not analyzed in any detail) by Conway and Kochen; the floor remains open for the latter.
2. Regarding the earlier work of Bell, Conway and Kochen (2006) acknowledged that:

> 'Our result is by no means the first in this direction. It makes use of the notorious quantum mechanical entanglement brought to light by Einstein, Podolsky, and Rosen, which has also been used in various forms by J.S. Bell, Kochen and Specker, and others to produce no-go theorems that dispose of the most plausible hidden variable theories. Our theorem seems to be the strongest and most precise result of this type.'

The precise relationship between the FWT and Bell's Theorem was analyzed in detail in Cator and Landsman (2014), with the following conclusion:

(a) Both Bell (1964) and the original version of the FWT in Conway and Kochen (2009) use an informal way of talking about free settings, granting which both establish a contradiction between determinism, locality (in the sense of Bell, which in the presence of determinism reduces to what is called parameter independence), and quantum mechanics. The difference lies in three facts:
   i. Bell relies on probability theory (whereas the FWT does not).
   ii. The (optical) corner of quantum mechanics used in Bell's Theorem may be replaced by the corresponding experimental results, whereas the FWT uses uncontroversial yet untested predictions about massive spin-1 particles.
   iii. The FWT must assume perfect (EPR) correlations, which are difficult to realize and hence are avoided by later versions of Bell's Theorem (i.e. through the well-known CHSH inequalities rather than the original Bell inequalities).

(b) The same three differences persist also in the new versions of both Bell's Theorem and the FWT proposed by Cator and Landsman (2014), in which the experimentalists' "freedom" of choosing settings is defined rigorously (in a probabilistic and a deterministic framework, respectively).

3. Conway and Kochen (2006) themselves already anticipated the criticism of circularity on the very first page of their first paper:

> "'I saw you put the fish in!" said a simpleton to an angler who had used a minnow to catch a bass. Our reply to an analogous objection would be that we use only a minuscule amount of human free will to deduce free will not only of the particles inside ourselves, but all over the universe.'

This did not stop Wüthrich (2011) from concluding that:

> 'Their case against determinism thus has all the virtues of theft over honest toil. It is truly indeterminism in, indeterminism out.'

Both are right: the FWT is far from circular, but its conclusion would be much more transparent if Wüthrich's charge could be dispelled. This is exactly what has been achieved in Cator and Landsman (2014), at least mathematically: we show that rather than "*indeterminism in*, *indeterminism out*", the thrust of the FWT is really: "*determinism in*, *constraints on determinism out*".

What is missing, then, from both the original papers by Conway and Kochen (2006, 2009) and the reformulation of the FWT by Cator and the author, is a serious analysis of the (philosophical) kind of free will assumed in the theorem, and thence of the implications of the theorem for that specific kind. The present paper attempts to fill this gap. In fact, it bridges a canyon, in relating the philosophical prose typical of at least the Lewisian corner of the free will literature (which I briefly review in Section 2) to elementary mathematics of the kind relevant to the FWT. This is done in Section 3, upon which the actual application to the FWT in Section 4 is straightforward. Finally, Section 5 contains my conclusions.[1]

2. Compatibilist free will à la David Lewis

The first question is which philosophical notion of free will is

[1] Since I base my analysis on our own revised FWT, any conclusions from this analysis about the original version can only be indirect, but in my opinion the potential link between indeterminism in the quantum world and free will in humans is so feeble that even if we grant that the original FWT is non-circular (in that it proves such indeterminism, as claimed by Conway and Kochen), its implications for free will are at best speculative.

among the assumptions of our revised FWT. Determinism being among the other assumptions, the answer must surely lie in the compatibilist direction, and I argue that more specifically one is close to the well-known 'local miracle' variant thereof due to Lewis (1981).[2] Following Moore (1912, Chapter 6) and many others in his wake, Lewis attempts to make sense of the intuition that even in a deterministic world one in principle has the ability to act differently from the way one actually does, despite the fact that the latter was predetermined. A standard example is that Alice's hand still rests on the table, although she was able to raise it. According to Lewis (1981), the way out is to realize that there is a fundamental difference between:

- I am able to do something such that, if I did it, a law would be broken.
- I am able to break a law.

Namely, the latter is impossible (which seems uncontroversial) but the former is not (which needs explanation). The first possibility rests on the theory of counterfactuals of Lewis himself (1973, 1979), according to which the antecedent 'if I did it' leads me to consider the possible world in which doing 'something' is actually possible, whilst as many other things as possible are kept the same as in the actual world (the precise underlying measure of similarity is not important here). However—and this a potential source of confusion—the conclusion that 'a law would be broken' refers to the *actual* world: the world in which 'I did it' is a different one. Thus although Lewis's position is often called *local miracle compatbilism*, a miracle takes place neither in the actual world where Alice's hand is at rest, nor in the possible world where she raises it. In other words, a law is broken neither in the former nor in the latter world. As Beebee (2013, p. 62) explains:

> 'This is what Lewis means by a 'miracle': an event M is a miracle if and only if M occurs at *possible world* w, and M is contrary to some *actual* law (or combination of laws) L. The point here is that while M is a miracle in Lewis's sense, it is not contrary to any of w's laws of nature. At w, L simply isn't a law in the first place. So, as things *actually* happened—in the *actual* world—L is a law, and M does not occur, so there is no miracle in the usual sense of 'miracle'. M is only a 'miracle' in Lewis's special sense of 'miracle': something (M) happens in w that is contrary to the laws of nature in the *actual* world.'

Throughout his analysis, Lewis himself takes a "miracle" to be a modification of the *laws of nature* between different worlds. Alternatively, the difference between the actual world in which Alice rests her hand and the possible world in which she raises it might be attributed to a modification of the *state* of the world.[3] Indeed, Vihvelin (2013, pp. 164–165) allows this possibility in her unfolding of Lewis's first bullet point as the following disjunction:

1. *Slightly different past*: If I had raised my hand, the past would still have been exactly the same until shortly before the time of my decision.
2. *Slightly different laws:* If I had raised my hand, the laws would have been ever so slightly different in a way that permitted a divergence from the lawful course of actual history shortly before the time of my decision.

Lewis (1981) himself rejects the first possibility, referring to his (1979) for an explanation:

> 'the way the future is depends counterfactually on the way the present is (…) [much as] the present depends counterfactually on the past (…) [but] not so in reverse (…) we ordinarily assume that facts about earlier times are counterfactually independent of the supposition and so may freely be used as auxiliary premises.' (Lewis, 1979, pp. 455–456).

Lewis (1979) proposes the hypothesis of *Asymmetry of Counterfactual Dependence*, suggesting that 'the mysterious asymmetry between open future and fixed past is nothing else than the asymmetry of counterfactual dependence' (p. 462). If this hypothesis is true, it would add a "philosophical" direction of time to the "physical" ones that have been proposed since the time of Boltzmann (Callender, 2011; Price, 1996; Zeh, 2007). From the point of view of determinism as commonly understood in physics (Earman, 2007), however, the laws of nature are time-symmetric.[4] In view of this, unlike Lewis himself I am open to the second way in which Alice could (counterfactually) have raised here hand, namely through an instant (counterfactual) modification of the state of the world, as in Bennett (1984). This option has been further explicated by Vihvelin (2013, p. 165) as:

1. *Same laws:* If I had raised my hand, the laws would still have been the same.
2. *Completely different past:* If I had raised my hand, past history would have been different all the way back to the Big Bang.

Here I would prefer to write *Different Past*, since even though in this scenario the state indeed (by determinism) would have been different all the way back to the Big Bang, the entire trajectory of the world may or may not be close to the actual one, in some suitable metric. In this scenario, the two cases Lewis distinguishes take the following form:

- I am able to do something such that, if I did it, the state of the actual world at some earlier time would have been different.
- I am able to change the state of the actual world at some earlier time.

The latter remains impossible (like breaking a law), whereas it is the former that gives rise to free will.[5] Had he preferred this second scenario, Lewis (1981) would have been entitled *Are we free to change the states?* instead of *Are we free to break the laws?*

Thus there are two different ways in which possible world w (in which Alice lifts her hand) could differ from the actual world (in which her hand is at rest): either there is some minor modification in the law governing the time-evolution of the universe (typically given by a small modification of the Hamiltonian local to Alice), or, with fixed time-evolution, there is a minor modification of the state of the world (again typically local to Alice at least at the time she thinks about what to do with her hand). Both would be acceptable to modern physics and of course, a mixture of these possibilities might be considered, too.

[2] van Inwagen (2008) states that Lewis's paper is 'the finest essay that has ever been written in defense of compatibilism—possibly the finest essay that has ever been written about any aspect of the free will problem'. For a recent introductory survey of the field see Beebee (2013). Versions of compatibilism vastly different from Lewis's have recently been proposed by Berofsky (2012) and Ismael (2016).

[3] In the philosophical literature this is called *backtracking*, cf. e.g. Lewis (1979) and Kapitan (2002).

[4] At least in so far as they apply to everyday situations relevant to the free will debate, hence excepting some eccentric corners of elementary particle physics involving *CP*-violation, accessible only in accelerators.

[5] As noted by Vihvelin (2013, Section 6.2), this still suffices to undermine the Consequence Argument.

3. Compatibilist free will revisited

Either way, as a considerable literature suggests (e.g., Beebee, 2003; Berofsky, 2012; Dorr, 2016; Fischer, 1994; Graham, 2008; Ismael, 2016; Oakley, 2006; Pendergraft, 2011), the tension between determinism and freedom remains worrying. To relax it, I present a simple mathematical framework that captures the spirit of compatibilist free will à la Lewis, including the idea of *agency* (which is an important aspect of free will). Here the intuition is that free will involves a separation between the agent, Alice, (who is to exercise it) and the rest of the world, under whose influence she acts: think of Alice as a deterministic chess computer, and of the rest of the world as her opponent. Her moves are determined by those of her opponent as well as by her own deliberations; the idea of free will then firstly means that different programs would play differently against the same opponent, and secondly—in the spirit of Lewis—that Alice has the ability to play differently from the way she actually does. Through this separation, I attempt to make the notion of other worlds close to the actual one precise (in a way that differs from Lewis).

I admit that what follows is a rather simple-minded idealization of any notion of (compatibilist) free will relevant to human behaviour in general (not to speak of moral issues related to free will), but I do believe that it is an appropriate approach to Alice's (alleged) free will in choosing the settings of her experiment (see also the Conclusion).

Let X be the state space of the world, and let

$$I\colon X \to X_I; \tag{3.1}$$

$$O\colon X \to X_O, \tag{3.2}$$

be variables that describe the agent and the rest of the world, respectively; here X_I is some set of "inner states" of the agent, whereas X_O consists of "outer states" of the world.[6] Let X_A be some set whose elements denote possible actions a of the agent, and let

$$A\colon X \to X_A \tag{3.3}$$

be the function that describes which action

$$a = A(x) \in X_A \tag{3.4}$$

the agent takes if the world is in state x. I assume that

$$A = A(O, I), \tag{3.5}$$

in the sense that each action $a = A(o, i)$ is a response to both the outer state

$$o = O(x) \tag{3.6}$$

of the rest of the universe (perhaps restricted to some relevant part) and the inner state

$$i = I(x) \tag{3.7}$$

of the agent. More precisely, there exists some function

$$\hat{A}\colon X_O \times X_I \to X_A, \tag{3.8}$$

such that for each $x \in X$ one has

$$A(x) = \hat{A}(O(x), I(x)). \tag{3.9}$$

Determinism of Alice's behaviour, briefly called *Determinism* in what follows, is expressed by the above framework, combined with the usual assumption of (Laplacian) determinism,[7] stating that there is a dynamical law $\varphi\colon X \times \mathbb{R} \to X$, satisfying $\varphi(x, 0) = x$ and $\varphi(\varphi(x, s), t) = \varphi(x, s + t)$. Thus Alice's behaviour a at time t is determined by the state x_0 of the world at any time t_0 in the past (or future) through (3.9) with

$$x = \varphi(x_0, t - t_0). \tag{3.10}$$

Freedom is our second fundamental assumption underpinning compatibilist free will. It states that O and I are *independent* (or that o and i are free variables), in the sense that for each $(o, i) \in X_O \times X_I$ there is $x \in X$ for which (3.6) and (3.7) hold, i.e., the following function is surjective:

$$O \times I\colon X \to X_O \times X_I$$
$$x \mapsto (O(x), I(x)). \tag{3.11}$$

Rephrasing our earlier analysis in this elementary mathematical language, Lewis wants to make sense of the idea that although (according to determinism) Alice's action $a = \hat{A}(o, i)$ at some fixed time t is determined by the state x of the world at that time through (3.9) and hence through (3.10) it was determined also by any earlier state x_0 of the world at time t_0,[8] nonetheless, Alice was able to act otherwise at time t, e.g. she was able to do

$$a' = \hat{A}(o', i'), \tag{3.12}$$

but did not do so, because doing a' would illegally have modified the state x.

Alice's ability to do a' means that there exists a state x' of the world close to x in that

$$o' = O(x') = O(x) = o, \tag{3.13}$$

making the outer environment in which Alice acts the same as in the actual world, but

$$i' = I(x') \neq I(x) = i, \tag{3.14}$$

where i' should be similar to i in some appropriate sense, such that (3.12) holds.

The point, then, is that according to our *Freedom* assumption, there indeed *is* such a state x', for any given i' and (o,i). Thus the freedom the agent has is precisely what I have formalized as *Freedom*: even *given* the state o of the external influences on her behaviour (and possibly even the state of the rest of the world), there is a different admissible state x' of the world such that, had this state been actual, the agent would have done a' (although she in fact did a). Since the actual world at time t resides in state x, the state x' (at the same time) pertains to a possible world w different from the actual. The difference between the two scenarios discussed in the previous section just lies in the story of how w was led to state x': either the law φ (Lewis) or the state (Bennett) was modified ever so slightly prior to time t.

[6] A special case one may have in mind here is $X = X_I \times X_O$, where I and O are projections onto the first and the second factor, respectively. However, in general X_I and X_O need not exhaust X.

[7] Compare with a typical philosopher's definition, e.g.: 'determinism is the thesis that for every instant of time t, there is a proposition that expresses the state of the world at that instant, and if P and Q are any proposition that express the state of the world at some instants, then the conjunction of P together with the laws of nature entails Q' (Vihvelin, 2013, p. 3). Thus P and Q correspond to our states x at different times, and 'the laws of nature' are combined into our function φ. What I have added to this in my definition is that the state x not only determines future (and past) states, but also controls what is going on, such as Alice's actions, namely through functions like A. Without these, states mean little.

[8] This is the whole point of the so-called *Consequence Argument* Lewis challenges.

4. The Free Will Theorem revisited

In this section I argue that in the context of the Free Will Theorem, the freedom of choosing settings for experiments which Conway and Kochen attribute to physicists is precisely of the above compatibilist nature. The theorem then establishes a contradiction between the physics assumptions and the compatibilist free will assumption, so that (accepting just a small but fundamental corner of modern physics), the latter must fall.

The setting of the Free Will Theorem of Conway and Kochen (2009) was introduced by EPR (Einstein, Podolsky, & Rosen, 1935) and further studied by Bell (1964) and others. In current jargon two physicists, called Alice and Bob, are far apart whilst performing simultaneous experiments on some correlated two-particle state (their measurements need to be *spacelike separated*). In the situation considered by EPR each particle had a spatial degree of freedom and hence required an infinite-dimensional Hilbert space for its description, but the thrust of the argument comes out more clearly if each particle merely has an internal degree of freedom (and is "frozen" otherwise). Bell (1964) considered a pair of spin-$\frac{1}{2}$ particles, each of which has Hilbert space $\mathbb{C}^2$, but because of its reliance on the Kochen–Specker Theorem (which fails for $\mathbb{C}^2$) the Free Will Theorem requires one dimension more, i.e., $H = \mathbb{C}^3$, seen as the state space of a massive spin-1 particle.[9]

For my purposes, all we need to know is that the experiment has *settings* (a,b) "freely chosen" by Alice and Bob, respectively,[10] and *outcomes* in the form of triples λ, since in fact Alice and Bob each perform three simultaneous measurements.[11] Defining

$$\Lambda = \{(0, 1, 1), (1, 0, 1), (1, 1, 0)\}. \tag{4.1}$$

quantum mechanics prescribes that each outcome λ of such a triple measurement must lie in Λ (so that, in particular, each single measurement of the three performed by both Alice and Bob must have outcome 0 or 1). Furthermore, on a specific choice of the correlated two-particle state, namely

$$\psi_0 = (\vec{e}_1 \otimes \vec{e}_1 + \vec{e}_2 \otimes \vec{e}_2 + \vec{e}_3 \otimes \vec{e}_3)/\sqrt{3}, \tag{4.2}$$

where $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ is the standard basis of $\mathbb{C}^3$, quantum mechanics predicts *perfect correlation* of measurement outcomes in case that the settings of Alice and Bob agree.[12]

The Free Will Theorem of Conway and Kochen (2009) states that three assumptions called SPIN, TWIN, and MIN, which I will recall below, imply that the response of a spin-one particle to the kind of experiment described below 'is not a function of properties of that part of the universe that is earlier than this response (…).' This formulation contains an implicit assumption of determinism, whose precise nature only becomes clear from their proof, and which is akin to our formulation below, except for the crucial difference that the function they allude to only acts on the particle variables and not on the settings of the experiment, of which Conway and Kochen just say that the experimenters can 'freely choose' them. As explained in Section 1, I (and others) consider this unsatisfactory.

I therefore first state a revised set of assumptions.

- *Determinism of measurement outcomes*, briefly called *Determinism* in what follows, means that there is a state space X with associated functions

$$A: X \to X_A, \quad B: X \to X_B, \quad F: X \to \Lambda, \quad G: X \to \Lambda, \quad Z: X \to X_Z, \tag{4.3}$$

  where $X_A = X_B$ is the set of settings available to Alice and Bob and Λ is some set of possible outcomes, which completely describe the experiment, in the sense that each state $x \in X$ determines both its settings $(a = A(x), b = B(x))$ and its outcome (both in Alice's lab and in Bob's lab) $(\lambda = F(x), \gamma = G(x))$.
  The function Z describes all relevant physical variables except Alice and Bob,[13] and should also be chosen (by the theory in question) in such a way that

$$F = F(A, B, Z); \quad G = G(A, B, Z). \tag{4.4}$$

  More precisely, there exist certain functions

$$\hat{F}: X_A \times X_B \times X_Z \to \Lambda; \quad \hat{G}: X_A \times X_B \times X_Z \to \Lambda, \tag{4.5}$$

  such that for each $x \in X$ one has

$$F(x) = \hat{F}(A(x), B(x), Z(x)); \quad G(x) = \hat{G}(A(x), B(x), Z(x)). \tag{4.6}$$

  As in the previous section, these rules should be supplemented with Laplacian determinism in order to deserve the name "Determinism" and have the right interpretation, but the above is what is needed in the proof of the Free Will Theorem.
- *Freedom* then states that A, B, and Z are *independent* in that for each $(a, b, z) \in X_A \times X_B \times X_Z$ there is an $x \in X$ for which $A(x) = a$, $B(x) = b$, and $Z(x) = z$.
- *Nature* requires that Λ is given by (4.1), and that perfect correlation holds.
- *Context locality* states that $F(A, B, Z)$ is independent of B and that $G(A, B, Z)$ is independent of A.[14] In other words, sharpening (4.4), we actually have

$$F = F(A, Z); \quad G = G(B, Z). \tag{4.7}$$

We then have the following version of the Free Will Theorem (Cator & Landsman, 2014):

> *Determinism, Freedom, Nature, and Context locality are contradictory.*

I refer to the paper just cited for the proof; after some new initial steps, the argument quickly reduces to the one due to Conway and Kochen (2009), whose assumptions are a subset of ours: their SPIN and TWIN are the first and second half of our *Nature* axiom, whilst their MIN expresses a form of context locality as well as the loose assumption that Alice and Bob may 'freely choose' their settings a and b, respectively. Accordingly, in terms of our notation, Conway and Kochen only use the parameter space Z, rather than the full state space X we need in order to consistently axiomatize determinism.[15]

[9] The price of this extra dimension is that the experiment whose outcome provides the empirical input for the Free Will Theorem has not actually been performed, except at a single wing (Huang, Li, Zhang, Pan, & Guo, 2003). However, the predictions of quantum mechanics are uncontroversial and will serve as input instead.

[10] In the context of the FWT the settings are frames, i.e., orthonormal bases of $\mathbb{R}^3$ defined up to a sign.

[11] If $a = (\vec{u}_1, \vec{u}_2, \vec{u}_3)$ is the basis chosen by Alice, she measures the three observables $\left(J^2_{\vec{u}_1}, J^2_{\vec{u}_2}, J^2_{\vec{u}_3}\right)$, where $J_{\vec{u}}$ is the angular momentum of the particle along the unit vector $\vec{u}$. Likewise for Bob.

[12] Writing $b = (\vec{v}_1, \vec{v}_2, \vec{v}_3)$ for Bob's basis, perfect correlation here means that if $\vec{u}_i = \vec{v}_j$, then the measured value of Alice's observable $J^2_{\vec{u}_i}$, which is 0 or 1, always coincides with Bob's $J^2_{\vec{v}_i}$.

[13] The parameter space X_Z includes the state of the pair of particles but its precise form is irrelevant.

[14] The name *Context locality*, which I learnt from M. Seevinck, distinguishes this condition from various other notions of locality such as *Einstein locality* (Haag, 1992), which is satisfied in quantum field theory, or *Bell locality* (Bell, 1990), which quantum mechanics violates. There is a probabilistic version of Context locality, called *No signalling* (or, for hidden variable theories, *Parameter Independence*), which holds in quantum mechanics. Context locality seems the weakest locality condition one may reasonably impose.

[15] In any case, the essence of the proof lies in the argument that perfect correlation together with context locality implies non-contextuality.

5. Conclusion

On my technical (re)formulations of both local miracle compatibilism (cf. Section 3) and the Free Will Theorem of Conway and Kochen (in Section 4), the freedom Alice and Bob enjoy in choosing their settings is precisely an instance of the kind of free will proposed by Lewis (1981). More precisely, the mathematical analogy is between:

- The triple $(o, i, a) \in X_O \times X_I \times X_A$ in the philosophical analysis of Alice's "free" choice of either resting (a) or raising (a') her hand —now seen as her choice for either the actual setting of her experiment or some other—as determined by the outer state o of the world and her own inner state i. Thus a is her actual setting, predicated on the state $x \in X$ and hence on $(o, i) = (O(x), I(x))$, which finally yield $a = \hat{A}(o, i)$.
- The triple $(a, z, \lambda) \in X_A \times Z \times \Lambda$ in the experimental setting of the FWT, where a is the setting of Alice's wing of the experiment (which from the perspective of the spin-1 particle plays the role of the outer state of the world), z is the inner state of the particle, and λ is the outcome of Alice's measurement.

In the spirit of Conway and Kochen, in the above analogy the Alice of the first bullet point (whose "free will" they after all believe to be ultimately a consequence of the "free choice" of elementary particles, cf. Section 1) plays the role of the spin-1 particles in the second bullet point. Conversely, the Alice of the second point is like the observable O in the first, which determines the external situation to which our particle responds (namely the setting a of the experiment, whose role in the first point is played by the outer state o of the world, to which Alice responds). Thus the analogy holds both mathematically and conceptually.

Granting this analogy, the Free Will Theorem establishes a contradiction between:

- the physics assumptions, i.e., *Nature* and *Context locality*;
- the compatibilist free will assumption, i.e., *Determinism* and *Freedom*.

Accepting the former, the latter must fall. Making this choice, one should realize that the physics assumptions on the one hand form just a small corner of modern physics (from which point of view they are weak), but on the other hand have singled out the corner in which the two fundamental theories of quantum mechanics and special relativity meet and are brought to a head (from which perspective they are strong). Either way, despite the lack of explicit experimental backing (which is available for Bell's Theorem),[16] few people would be willing to reject these physics assumptions (and in any case the exercise is to determine what modern physics has to say about free will, which presupposes the former).

Finally, although the intention of Conway and Kochen was to *support* unspecified versions of libertarian free will through modern physics, our reformulation of their theorem (which removes the threat of circularity) gives a more subtle picture: the FWT (revisited) challenges one particular version of *compatibilist free will*. As such, it only provides indirect support for *libertarian free will*, namely by weakening one of its competitors.

Acknowledgement

The author is much indebted to Jeremy Butterfield, Ronnie Hermens, Gijs Leegwater, Fred Muller, Marc Slors, and two excellent anonymous referees for comments on drafts of this paper. Also many thanks to Helen Beebee for checking Section 2.

[16] The set of outcomes (4.1) at each wing has actually been experimentally vindicated (Huang et al., 2003), but the correlations between the outcomes of Alice and Bob have not been, so far.

References

Bassi, A., & Ghirardi, G. C. (2007). The Conway–Kochen argument and relativistic GRW models. *Foundations of Physics*, *37*, 169–185.
Beebee, H. (2003). Local miracle compatibilism. *Noûs*, *37*, 258–277.
Beebee, H. (2013). *Free will: An introduction*. New York: Palgrave MacMillan.
Bell, J. S. (1964). On the Einstein–Podolsky–Rosen Paradox. *Physics*, *1*, 195–200.
Bell, J. S. (1990). La nouvelle cuisine In: A. Sarlemijn, & P. Kroes (Eds.), *Between science and technology* (pp. 97–115). Amsterdam: Elsevier.
Bennett, J. (1984). Counterfactuals and temporal direction. *The Philosophical Review*, *93*, 57–91.
Berofsky, B. (2012). *Nature's challenge to Free Will*. Oxford: Oxford University Press.
Brown, H., & Svetlichny, G. (1990). Nonlocality and Gleason's lemma. Part I. Deterministic theories. *Foundations of Physics*, *20*, 1379–1387.
Callender, C. (2011). Thermodynamic asymmetry in time. *The stanford encyclopedia of philosophy (fall 2011 edition)*, In: E.N. Zalta, (Ed.), ⟨http://plato.stanford.edu/archives/fall2011/entries/time-thermo/⟩.
Cator, E., & Landsman, N. P. (2014). Constraints on determinism: Bell versus Conway–Kochen. *Foundations of Physics*, *44*, 781–791.
Clifton, R. (1993). Getting contextual and nonlocal elements-of-reality the easy way. *American Journal of Physics*, *61*, 443–447.
Conway, J. H., & Kochen, S. (2006). The Free Will Theorem. *Foundations of Physics*, *36*, 1441–1473.
Conway, J. H., & Kochen, S. (2009). The strong Free Will Theorem. *Notices of the American Mathematical Society*, *56*, 226–232.
Dorr, C. (2016). Against counterfactual miracles. *Philosophical Review*, *125*, 241–286.
Earman, J. (2007). Aspects of determinism in modern physics In: J. Butterfield, & J. Earman (Eds.), *Handbook of philosophy of science, philosophy of physics*, vol. 2 (pp. 1369–1434). Amsterdam: Elsevier.
Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, *47*, 777–780.
Fischer, J. M. (1994). *The Metaphysics of Free Will*. Oxford: Blackwell.
Graham, P. A. (2008). A defence of local miracle compatibilism. *Philosophical Studies*, *140*, 65–82.
Goldstein, S., Tausk, D. V., Tumulka, R., & Zanghi, N. (December 2010). What does the free will theorem actually prove? *Notices of the AMS*, 1451–1453.
Haag, R. (1992). *Local quantum physics: Fields, particles, algebras*. Heidelberg: Springer.
Hemmick, D. L., & Shakur, A. M. (2012). *Bell's theorem and quantum realism: Reassessment in the light of the Schrödinger Paradox*. Heidelberg: Springer.
Hermens, R. (2014). Conway–Kochen and the finite precision loophole. *Foundations of Physics*, *44*, 1038–1048.
Hermens, R. (2016). *Philosophy of quantum probability*. (Ph.D. thesis). University of Groningen.
Heywood, P., & Redhead, M. (1983). Nonlocality and the Kochen–Specker paradox. *Foundations of Physics*, *13*, 481–499.
't Hooft, G. (2007). *The free-will postulate in quantum mechanics*. arXiv:quant-ph/0701097.
Huang, Y.-F., Li, C.-F., Zhang, Y.-S., Pan, J.-W., & Guo, G.-C. (2003). Experimental test of the Kochen–Specker theorem with single photons. *Physical Review Letters*, *90*, 250401.
Ismael, J. T. (2016). *How physics makes us free*. Oxford: Oxford University Press.
Kapitan, T. (2002). A master argument for incompatibilsm? In: R. Kane (Ed.), *The oxford handbook of Free Will* (pp. 127–157). Oxford: Oxford University Press.
Lewis, D. (1973). *Counterfactuals*. Cambridge (MA): Harvard University Press.
Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*, 455–476.
Lewis, D. (1981). Are we free to break the laws? *Theoria*, *47*, 112–121.
Moore, G. E. (1912). *Ethics*. New York: Henry Holt and Company See also the free online version at ⟨http://fair-use.org/g-e-moore/ethics/⟩.
Oakley, S. (2006). Defending Lewis's local miracle compatibilism. *Philosophical Studies*, *130*, 337–349.
Pendergraft, G. (2011). The explanatory power of local miracle compatibilism. *Philosophical Studies*, *156*, 249–266.
Price, H. (1996). *Time's arrow and archimedes' point: New directions for the physics of time*. Oxford: Oxford University Press.
Roberts, S. (2015). *Genius at play: The curious mind of John Horton Conway*. New York: Bloomsbury.
Stairs, A. (1983). Quantum logic, realism, and value definiteness. *Philosophy of Science*, *50*, 578–602.
van Inwagen, P. (2008). How to think about the problem of free will. *Journal of Ethics*, *12*, 337–341.
Vihvelin, K. (2013). *Laws, and Free Will: Why determinism doesn't matter. Causes*. Oxford: Oxford University Press.
Wüthrich, C. (2011). Can the world be shown to be indeterministic after all? In: C. Beisbart, & S. Hartmann (Eds.), *Probabilities in physics* (pp. 365–389). Oxford: Oxford University Press.
Zeh, H. D. (2007). *The physical basis of the direction of time* (5th Edition). Berlin: Springer.