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
landsman@math.ru.nl

October 5, 2015

Abstract

The (Strong) Free Will Theorem of Conway & 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). Although Conway and Kochen informally claim that their theorem supports indeterminism and, in its wake, a libertarian agenda for free will, inferring the former from the Free Will Theorem is a *petitio principii* (as has been pointed out before). Of course, this also considerably weakens the case for the latter (at least from the theorem).

A recent reformulation of the Free Will Theorem due to Cator and the author (2014) uses a mathematically precise formulation of the idea that experimentalists freely choose their settings even in a deterministic world, and thence construes the theorem as providing constraints on determinism (as opposed to proving indeterminism). In the present paper we argue that the concept of free will used in this version of the theorem is essentially the one proposed by Lewis (1981), also known as 'local miracle compatibilism'. To do so, we give a mathematical reformulation of this idea, which we believe to be well within its spirit and which may be of independent interest.

On this double reformulation the Free Will Theorem shows that local miracle compatibilism contradicts the parts of modern physics used in its assumptions. Therefore, although the intention of Conway and Kochen was to support free will through their theorem, what they actually achieved is the opposite: a well-known, philosophically viable version of free will now turns out to be incompatible with physics!

Keywords: Free Will Theorem, Local Miracle Compatibilism

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

1 The Free Will Theorem

The Free Will Theorem (FWT) of Conway & 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.[1]

Conway and Kochen (2009) 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.'

One wonders if such far-reaching conclusions are warranted by the actual technical content of the Free Will Theorem. In particular, although the joint assumption of determinism and freedom (of the kind summarized above) is also made in the famous paper by Bell (1964),[2] this conjunction is especially uncomfortable in the setting of the FWT, whose very emphasis on free will in an attempt to refute determinism seems almost self-contradictory (though in fact it is not, as we will show). The authors already anticipate criticism of this kind on the very first page of their first paper (Conway & Kochen, 2006), writing:

> '"I saw you put the fish in!" said a simpleton to an angler who had used a minnow to catch a bass.'

Indeed, also after more serious philosophical analysis, Wüthrich (2011) concluded that:

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

Our formulation of the FWT (Cator & Landsman, 2014) avoids this criticism, as the original allusion to undefined free will in allowing arbitrary settings of the experiment has been replaced by explicit (super)determinism including the settings, to subsequently recover the situation of the original FWT through a natural independence assumption. Moreover, it has the advantage of clearly stating the assumptions on which the proof is based. Thus we show that rather than "*indeterminism in, indeterminism out*", the thrust of the Free Will Theorem is really: "*determinism in, constraints on determinism out*".

[1] 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). Analogous results had previously been published by Heywood & Redhead (1983), Stairs (1983), Brown & Svetlichny (1990), and Clifton (1993), of which only the first paper was cited by Conway and Kochen. Moreover, the close relationship to Bell (1964), whose assumptions and conclusions are quite similar to those of the FWT (cf. Bassi & Ghirardi, 2007; Cator & Landsman, 2014), might have been acknowledged in their papers. Indeed, the main novelty of the FWT lies in the emphasis Conway and Kochen (unlike the other authors mentioned in this footnote) put on free will. We also refer to Goldstein et al (2010) for a critique of an extension of the FWT to stochastic theories (by which Conway and Kochen apparently overplayed their hand).

[2] In whose context the tension between causality and freedom is discussed in Seevinck & Uffink (2011).

2 Compatibilist free will à la David Lewis

The question arises which philosophical notion of free will is among the assumptions of the (thus revised) FWT.[3] Determinism being among the other assumptions, the answer must surely lie in the compatibilist direction. Indeed, we argue that more specifically we are close to the well-known 'local miracle' variant thereof due to Lewis (1981).[4] Following Moore (1912, Ch. 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. However, she could not have raised it with a speed greater than that of light, so her ability remains constrained by the laws of nature. In Lewis's phrasing (which, as we shall see, we regard as somewhat misleading), 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.

According to Lewis, 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:[5]

[3] Though taking a different turn, our analysis below was significantly influenced by 't Hooft (2007).

[4] It has been said 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' (van Inwagen, 2008). For a recent introductory survey of the field see Beebee (2013).

[5] Analogously in Beebee (2003, pp. 261–262). Unfortunately, confusion may arise if the above quotation 'if I did it, a law would be broken' from Lewis (1981) is subjected to the following explanation:

> 'On Lewis's account of counterfactuals, the *truth conditions* for counterfactuals—what makes them true—are as follows. Suppose we have the counterfactual 'if A had been the case, B would have been the case' (so if A is 'I miss the bus' and B is 'I'm late', this counterfactual just says, 'if I'd missed the bus, I would have been late'). This counterfactual will be true if and only if, *at the closest possible world to the actual world* at which A is true, B is also true. So, our sample counterfactual, 'if I'd missed the bus, I would have been late', is true if and only if: *at the closest possible world to the actual world* at which I miss the bus, I'm late.' (Beebee, 2013, p. 60).

Removing any possible remaining doubt, on p. 62 she mentions that the closest possible world where I miss the bus is the world w. According to this explanation, then, Lewis's sentence 'if I did it, a law would be broken', would mean that *at the closest possible world to the actual world* in which I did it, a law *is* broken, i.e., in w. But according to Beebee's definition quoted in the main text of what Lewis means by a miracle, apparently this is not the right reading (and indeed it would, in our view, be nonsensical). Moreover, Lewis (1981) emphasizes that in the first bullet point in the main text above—which he defends—it is not the agent who would break a law, whereas in the second bullet point —rejected by Lewis—it is; in the first it is the breaking of some law at an earlier time that enables the agent to do what she, in our actual world, did not do. We stick to the interpretation in our main text, regarding Lewis's phrasing as awkward.

> '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.'

Vihvelin (2013, pp. 164–165) states Lewis's first bullet point as the following conjunction:

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.

This way of paraphrasing Lewis leads to a point that merits discussion. Throughout his analysis, Lewis takes a "miracle" to be a change in the laws of nature. 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 change in the *state* of the world.[6] Lewis (1981) rejects this possibility, referring to his (1979) for an explanation. Here we read:

> '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) therefore 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, cf. Price (1996), Zeh (2007), and Callender (2011). From the point of view of determinism as commonly understood in physics (Earman, 2007), however, at least the laws of nature are time-symmetric.[7] In view of this, unlike Lewis himself we are open to the second way in which Alice could (counterfactually) have raised here hand, namely through an instant (counterfactual) change in the state of the world, as in Bennett (1984). This has been explicated by Vihvelin (2013, p. 165), too:

1. **Same Laws:** If I had raised my hand, the laws would still have been exactly 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.

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

[7] 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.

Here we would actually prefer to write **Different Past**, since even though in this scenario the state indeed (by determinism rephrased as the uniqueness of the solution of the equation of motion for the state of the universe with given value at any fixed time) 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.[8] 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 change in the law governing the time-evolution of the universe (typically given by a small change in the Hamiltonian local to Alice), or, with fixed time-evolution, there is a minor change in 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

3 Compatibilist free will revisited

Either way, as a considerable literature suggests (e.g., Fischer, 1994; Beebee, 2003; Oakley, 2006; Graham, 2008; Pendergraft, 2011), the tension between determinism and freedom remains worrying. To relax it, we we will now 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. Namely, let X be the state space of the world, and let

$$I : X \to X_I; \tag{3.1}$$
$$O : X \to X_O \tag{3.2}$$

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.[9] Let X_A be some set whose elements denote possible actions a of the agent, and let

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

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. We assume that

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

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

[9] 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.

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

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

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,[10] stating that there is a dynamical law $\varphi : X \times \mathbb{R} \to X$, 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:

$$\begin{aligned} O \times I : X &\to X_O \times X_I \\ x &\mapsto (O(x), I(x)). \end{aligned} \tag{3.11}$$

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—this is the whole point of the so-called *Consequence Argument* Lewis challenges—through (3.10) it was determined also by any earlier state x_0 of the world at time t_0;

- 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 changed the state x.

[10] 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 we have added to this in our 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 nothing.

Alice's ability to do a' now simply means that there exists a state x' of the world close to x in the sense 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 as a slight change in the state of Alice's brain), 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 we 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, necessarily, 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) changed ever so slightly prior to time t.

4 The Free Will Theorem revisited

In this section we show 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 & Kochen (2009) was introduced by EPR (Einstein, Podolsky, and 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.[11]

For our purposes, all we need to know is that the experiment has *settings* (a, b) "freely chosen" by Alice and Bob, respectively,[12] and *outcomes* in the form of triples λ, since in fact Alice and Bob each perform three simultaneous measurements.[13] Defining

$$\Lambda = \{(0, 1, 1), (1, 0, 1), (1, 1, 0)\}. \tag{4.15}$$

[11]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 et al, 2003). However, the predictions of quantum mechanics are uncontroversial and will serve as input instead.

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

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

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,[14] quantum mechanics predicts *perfect correlation* of measurement outcomes in case that the settings of Alice and Bob agree.[15]

The Free Will Theorem of Conway and Kochen (2009) states that three assumptions called SPIN, TWIN, and MIN, which we 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 §1, we (and others) consider this unsatisfactory.

We 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

  $$A : X \to X_A,\ B : X \to X_B,\ F : X \to \Lambda,\ G : X \to \Lambda,\ Z : X \to X_Z, \tag{4.16}$$

  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

  $$(a = A(x), b = B(x)), \tag{4.17}$$

  and its outcome (both in Alice's lab and in Bob's lab)

  $$(\lambda = F(x), \gamma = G(x)). \tag{4.18}$$

  The function Z describes all relevant physical variables except Alice and Bob,[16] and should also be chosen (by the theory in question) in such a way that

  $$F = F(A, B, Z); \tag{4.19}$$
  $$G = G(A, B, Z). \tag{4.20}$$

  More precisely, there exist certain functions

  $$\hat{F} : X_A \times X_B \times X_Z \to \Lambda; \tag{4.21}$$
  $$\hat{G} : X_A \times X_B \times X_Z \to \Lambda, \tag{4.22}$$

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

  $$F(x) = \hat{F}(A(x), B(x), Z(x)); \tag{4.23}$$
  $$G(x) = \hat{G}(A(x), B(x), Z(x)). \tag{4.24}$$

  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.

---

[14] Namely, $\psi_0 = (\vec{e}_1 \otimes \vec{e}_1 + \vec{e}_2 \otimes \vec{e}_2 + \vec{e}_3 \otimes \vec{e}_3)/\sqrt{3}$, where $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ is the standard basis of $\mathbb{C}^3$.

[15] 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 $J^2_{\vec{v}_i}$.

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

- ***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 (4.15), and that perfect correlation holds.
- ***Locality*** makes $F(A, B, Z)$ independent of B and $G(A, B, Z)$ independent of A.[17]

  In other words, sharpening (4.19) - (4.20), we actually have

$$F = F(A, Z); \tag{4.25}$$
$$G = G(B, Z). \tag{4.26}$$

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

Determinism, Freedom, Nature, and Locality are mutually contradictory.

We refer to the paper just cited for the proof; after some new initial steps, the argument quickly reduces to the one due to Conway & 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.[18]

5 Conclusion

Our technical (re)formulations of both local miracle compatibilism (cf. §2) and the Free Will Theorem of Conway and Kochen (in §4) show that 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"

[17]A more precise name for this condition would be *Context locality*, distinguishing it 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 Contextlocality, 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.

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

of elementary particles, cf. §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 our analogy holds both mathematically and conceptually.

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

- the physics assumptions, i.e., *Nature* and *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),[19] 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).

Thus the irony is that although the outspoken intention of Conway and Kochen was to *support* free will through modern physics (and elementary logic), what they actually achieved is the opposite: one of the few philosophically viable versions of free will (Vihvelin, 2013) has now been shown to be *incompatible* with quantum theory and relativity.[20]

Although we find this conclusion worth pointing out, in all honesty it is hardly surprising: for according to Maudlin (2007), Lewis's "Humean supervenience" program, into which his theory of free will seems firmly embedded, is at odds with modern physics anyway, especially with the quantum-mechanical assumption in the FWT. So we agree.

[19] The set of outcomes (4.15) 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.

[20] To their defence, it is fair to say that Conway and Kochen never intended to support compatibilist free will in the first place. It is hard to find more scathing comments on compatibilism than the following ones by Conway: 'Compatibilism in my view is silly. Sorry, I shouldn't just say straight off that it is silly. 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 nobody would *ever* have come up with compatibilism if they thought, as turns out to be the case, that science wasn't deterministic. The whole business of compatibilism was to reconcile what science told you at the time, centuries ago down to 1 century ago: Science appeared to be totally deterministic, and how can we reconcile that with free will, which is not deterministic? So compatibilism, I see it as out of date, really. It's doing something that doesn't need to be done. However, compatibilism hasn't gone out of date, certainly, as far as the philosophers are concerned. Lots of them are still very keen on it. How can I say it? If you do anything that seems impossible, you're quite proud when you appear to have succeeded. And so really the philosophers don't want to give up this notion of compatibilism because it seems to damned clever. But my view is it's really nonsense. And it's not necessary. So whether it actually is nonsense or not doesn't matter.' (Conway, quoted in Roberts, 2015, pp. 361–362).

Or, in rather more academic parlance: '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).

Acknowledgement

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

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. *Between Science and Technology*, eds. Sarlemijn, A., Kroes, P., pp. 97–115. Amsterdam: Elsevier.

Bennett, J. (1984). Counterfactuals and temporal direction. *The Philosophical Review* 93, 57–91.

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)*, Zalta, E.N. (ed.). `http://plato.stanford.edu/archives/fall2011/entries/time-thermo/`.

Cator, E., Landsman, K. (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.

Earman J. (2007). Aspects of determinism in modern physics. *Handbook of Philosophy of Science, Vol. 2: Philosophy of Physics*, Butterfield, J., Earman, J., eds., 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. (2010). What does the free will theorem actually prove?. *Notices of the AMS* December 2010, 1451–1453.

Haag, R. (1992). *L*ocal Quantum Physics: Fields, Particles, Algebras. Heidelberg: Springer.

Hermens, R. (2015). *Philosophy of Quantum Probability.* PhD 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.

Inwagen, P. van (1975). The incompatibility of Free Will and Determinism. *Philosophical Studies* 27, 185–199. van Inwagen, 2008).

Inwagen, P. van (2008). How to think about the problem of free will. *Journal of Ethics* 12, 337–341.

Kapitan, T. (2002). A master argument for incompatibilsm? *The Oxford Handbook of Free Will*, ed. Kane, R., 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.

Maudlin, T. (2007). *The Metaphysics Within Physics*. Oxford: Oxford University Press.

Moore, G.E. (1912). *Ethics*. New York: Henry Holt and Company. See also 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.

Seevinck, M.P., Uffink, J. (2011). Not throwing out the baby with the bathwater: Bell's condition of local causality mathematically 'sharp and clean'. *Explanation, Prediction, and Confirmation* 2, 425–450.

Stairs, A. (1983). Quantum logic, realism, and value definiteness. *Philosophy of Science* 50, 578–602.

Vihvelin, K. (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. Oxford: Oxford University Press.

Wüthrich, C. (2011). Can the world be shown to be indeterministic after all? *Probabilities in Physics*, eds. Beisbart, C., Hartmann, S., pp. 365–389. Oxford: Oxford University Press.

Zeh, H.D. (2007). *The Physical Basis of the Direction of Time. 5th Edition*. Berlin: Springer.