# Free Will in the Many-Worlds Interpretation of Quantum Mechanics

David John Baker
djbaker@umich.edu

November 19, 2022

**Abstract**

David Wallace has argued that there is no special problem for free will in the many-worlds interpretation of quantum mechanics, beyond the well-known problem of reconciling free will with physical determinism. I argue to the contrary that, on the plausible and popular "deep self" approach to compatibilism, the many-worlds interpretation does face a special problem. It is not clear on the many-worlds picture how our actions can issue from our most central character traits, given that copies of us in other branches are certain to act differently than we do.

## 1 Introduction

The Everett or many-worlds interpretation of quantum theory has lately grown in popularity (Wallace, 2012, 423) and (thanks to recent work) in plausibility. The time is ripe, I think, to wrestle with the human implications of the interpretation. If Everettians are right, your life is not a straight line stretching from past to future. It is more like a tree, branching again and again into many lives–many people in other parallel worlds, each with an equal claim to descend from your past self. And whenever a quantum-mechanically random event occurs, a new person in a new world will see each and every possible outcome.

Plausibly, human decisions are random in the relevant sense. So typically, when you face a choice, your descendants in different branches will make and act out different decisions. The implications for free will are frightening, and potentially dehumanizing.

In his story "Singleton," Greg Egan (2002) portrays a scientist whose proudest moment is an act of heroism in which he ran to the aid of a man being brutally beaten. Later in life, he comes to believe the Everett interpretation. He realizes that if his understanding of physics is correct, many of his other selves in other worlds must have succumbed to fear and left the man to his attackers. Considering his daughter's future, he comes to believe that in an Everettian world it will be "impossible for her ever to act wholly in accordance with her ideals." (Egan, 2002) And this, he thinks, would rob her of free will.

In his exhaustive defense of the Everett interpretation, David Wallace considers Egan's fears about free will, and suggests they are misplaced. Granting for the sake of argument that our decisions will be different in different branches, he writes:

> Does this in any way undermine the fact that the choice (or the multiple choices, if the agent's decision-making process is quantum-mechanically random) is freely made? Only if there is a clash between the idea of a choice being *free* and the idea of it being the result of some mechanistic physical process (deterministic or otherwise). (Wallace, 2012, 136)

In other words, Wallace subsumes Egan's worry under the umbrella of incompatibilism.

This is unfair to Egan, I think. His concern is more sophisticated than the simple observation that our actions, in an Everett world, are determined by the laws of nature. "Free will," he writes, "was a slippery notion, but to me it simply meant that your choices were more or less consistent with your nature." (Egan, 2002) This sounds a lot like compatibilism. Even from within this compatibilist picture, Egan finds space to worry about free will– because in an Everett world, our choices are never *fully* consistent with our natures. We only seem to express ourselves in our decisions because our perceptions are confined to a single world.

I want to develop and defend Egan's worry about free will in the many-worlds interpretation, from a compatibilist point of view. There is much more to this worry than Wallace grants. For among the most promising versions of compatibilism is the *deep self view*, on which our actions are free when they express our most foundational traits of character. And if our universe is a branching collection of worlds as the Everett interpretation suggests, it is impossible for our actions to express our deep selves. Many-worlds quantum mechanics is therefore less hospitable to free will than even other deterministic physical theories.

Expounding on Egan's worry will require a more detailed picture of the many-worlds interpretation, and especially the attendant accounts of probability and personal identity.

This will be the task of Section 2. Then in the following section, I will present an Egan-inspired argument that deep-self compatibilism cannot succeed in an Everettian world.

## 2 Preliminaries

The present-day many-worlds interpretation is *fundamentally* a theory of a single universe. But it is a universe with a very rich emergent structure, which behaves (to a good approximation) like a collection of branching worlds. The worlds branch in the sense that, whenever a process analogous to the measurement of a microscopic system occurs in a given world, that world is quickly replaced with several different causally isolated worlds, one for every possible outcome of the measurement.

Note that this branching behavior is not confined to literal measurements with scientific instruments. Rather, it occurs whenever the precise state of a microscopic system like an atom or elementary particle becomes strongly correlated with the state of its macroscopic environment. This process is called *decoherence.*

It is an interesting question whether decoherence-caused branching will generally occur at the point of human decision-making, leading to different branches in which the human agent chooses different options. Branching is going on pretty much all the time in many-worlds, but it may be that microscopic variables almost never become correlated with the specific macroscopic variables that determine our actions. If they do not, there will be very many worlds branching off from any given decision, but in all the branches the agent will make the same decision. The other possibility is that microscopic variables may play a large role in determining human actions, in which case any given decision will result in branches where the agent makes different choices. If this latter possibility is actual, the interpretation must contend with Egan's worry about free will.

This question can only settled by the details of the physics, which are unknown and likely to remain so for a long time. Decisions will be different in different branches if the human mind turns out to be a *classically chaotic* system, one in which final outcomes are sensitive to very small differences in initial conditions (Wallace, 2012, 64-99). On the face of it, it's plausible that the brain's decision-making activities could turn out to be chaotic in this sense, but of course the matter has not been rigorously investigated and is unlikely to be settled anytime soon. I propose to assume for purposes of argument that our decisions do turn out differently in different branches, and determine the implications of this assumption for human free will.

## 2.1 Probability: determinism and uncertainty

The problem of probability is the most hotly debated puzzle of many-worlds, and as we'll see it is relevant to the question of free will. As Wallace (2003) and Greaves (2004) first emphasized, there are really two problems of probability for the interpretation. The proponents of many-worlds claim to have solved this problem (Wallace, 2003; Deutsch, 1999), although as we'll see in Sec. 2.3, their solution may be incomplete in some important respects.

The other, which does concern us, is the *coherence problem*: The predictions of quantum theory are probabilistic, but on the many-worlds interpretation the laws of nature are deterministic. The overall branching structure of the universe at any time is uniquely determined by its past state. The appearance of indeterminism is explained by the fact that the different possible outcomes of an initial measurement will in general appear in different branches of the future state. But how can it make sense to say that one of these branches is "more probable" than another, when from an objective point of view it is certain that every outcome will occur in some branch?

We can motivate this problem by considering a more familiar thought experiment–the human fission cases discussed by Wiggins (1967) and Parfit (1971). Think of a person who is, like an amoeba, about to split into two people who are copies of the original. (This is analogous to many-worlds branching, in which an initial state including one person is replaced by a state including multiple people, equally causally connected with the original person.) We don't normally think it makes sense to treat such fission as a probabilistic process. In particular there is no apparent way to make sense of the notion that it is more or less probable that the initial agent will become one of the post-fission agents, rather than becoming the other.

But then how can we make sense of probability in the branching universes of many-worlds? Shouldn't the probability of any possible outcome be one, since every outcome happens in some branch or other? There are three approaches to this puzzle: the Subjective Uncertainty (SU) approach of Wallace (2003, 2012), the Post-Measurement Uncertainty (PMU) approach of Vaidman (1998) and Sebens and Carroll (forthcoming), and the Objective Determinism (OD) approach of Greaves (2004).

The SU approach insists that, despite the deterministic physics of many-worlds, there is a clear sense in which it is reasonable to be uncertain about the outcome of a branching event. Several forms of this view have been proposed, but the clearest, due to Saunders and Wallace (2008), rests on David Lewis's account of personal identity. Lewis (1976) has suggested that a human being is best understood as a temporally extended object, a thing composed of instantaneous temporal parts he calls person-stages. On this view, it is possible for a person-stage to be part of two distinct human beings, as in the case of fission. When human fission occurs, on Lewis's view there were two people all along–it's just that these two people share all their properties in common up until the time of fission.

Saunders and Wallace accept this picture of personal identity, which entails that, in many-worlds prior to branching, there are many people sharing your current person-stage. For example, if you are preparing to measure a particle's spin, your current stage will belong to the life of a person who goes on to observe spin up, and also a person in another branch who goes on to observe spin down. It thus must make sense, Saunders and Wallace argue, to wonder which of these many people you are, to assign some degree of belief to the proposition that you are the person who will see spin up, and so on. The probabilistic predictions of quantum theory consist of objective facts about which degrees of belief you should assign to these "who am I?" propositions.

The concept of uncertainty at the heart of SU may seem confused to you. (If it does, you have my sympathies.) Others have attempted to make sense of many-worlds probability without Saunders and Wallace's unusual commitments. Vaidman (1998) suggests that post-measurement self-locating uncertainty can solve the coherence problem. It seems beyond question that if, for example, I've performed a measurement (thereby causing branching) and haven't yet looked at its result, I am in a position to wonder which branch I now inhabit. Am I in the spin up branch or the spin down branch? It would seem that I ought to assign some degree of belief to both possibilities. So perhaps probability in many-worlds is an objective fact about what my self-locating degrees of belief should be in the moments following a measurement. This is the PMU view of probability.

Although the PMU solution is conceptually clearer than SU, it's not clear that post-measurement probabilities can do all the work that probability appears to do in quantum theory. We may need a more robust notion of what it means, prior to measurement, for one result to possess some probability. For those who reject the SU picture but deny that the PMU picture can give the whole story, the objective determinism (OD) view may be the best bet.

The OD view accepts that the outcomes of experiments in many-worlds are not a matter of chance, and that there is no sense to be made of pre-measurement uncertainty. But we may still make sense of the decision-theoretic *weight* of future branches, in the following sense. If we imagine a classical human fission event, there ought to be some fact of the matter about how the initial agent should weigh the interests of the future people who will result from the fusion. For example, if we offer to reward one of the agent's successors and punish the other, the agent ought to have some preferences about whether to take this deal, which will depend on how much she values the well-being of one successor relative to the other.

In the basic fission example, it is intuitive that we ought to weight each successor's interests equally. In the case of many-worlds branching, it isn't possible to count branches and treat outcomes with more branches as more probable. The concept of "branch" is vague enough that there won't generally be a sharp matter of fact about how many branches exist. (How little interaction must there be between two portions of the state before we should treat them as separate worlds? Presumably there is no sharp line to draw here.)

Greaves (2004) argues, however, that it is still entirely coherent to suppose that there is some rational constraint on how we ought to weigh the interests of our successors in future branches, and the importance of events in those branches. Moreover, she claims that what we think of as probability in quantum mechanics is actually a measure of this decision-theoretic importance. So when a laboratory physicist says that a spin measurement is twice as likely to come out up as it is to come out down, on Greaves's OD view we should take this to mean that the branches where the measurement comes out spin-up are twice as important (for decision-making purposes) as the branches where it comes out spin-down. She compares this view to Parfit's claim that what we consider important about survival can, in cases like human fission, come in degrees. Greaves suggests that when decoherence causes our world to branch, what we ordinarily call quantum "probability" is a measure of the degree to which we are survived by the resulting people.

The relative strengths and weaknesses of these three views are of course debatable. So is the question of whether any of the three is ultimately satisfactory. For our purposes, what matters is that there are no other viable-looking ways to understand probability in many-worlds.

## 2.2 Setting aside the linguistic account

It is worth noting here that a fourth account of Everettian probability, a *linguistic* account, has been proposed by Wallace (2012, Ch. 7). I have not counted this view among the viable accounts of probability because I take it to be decisively refuted by the objections of Bacciagaluppi and Ismael (2015). Let me briefly elaborate.

Wallace argues that if, in fact, we live in an Everettian universe, an ordinary folk utterance of (for example) "I'm uncertain about whether it will rain tomorrow; the chances are 50-50" should be understood to mean something like, *It will rain in some future branches and not others, and the amplitude of the rain branches is equal to that of the no-rain branches.* A more intuitive semantics for such utterances will count much of our commonplace discourse as false, running afoul of the principle of interpretive charity.

This makes the coherence problem trivially easy to solve, since it entails that our preexisting concept of objective chance already refers to quantum amplitude. But as Bacciagaluppi and Ismael point out, this supposedly charitable interpretation gets many of our dispositions wrong. For it is highly plausible (verging on obvious) that typical competent speakers do not possess the disposition to use the word 'uncertainty' in cases of branching under conditions of ideal information. If you were to present the physical facts about branching to the average person and ask them whether their word 'uncertainty' applies whenever one is about to branch, they would doubtless respond, "No, apparently we've used the word in such situations in the past, but we were mistaken about the facts. Now that we know the physics, we know that the events we would've previously said we were 'uncertain' about will inevitably occur in some branch or other. So 'uncertainty' is really the wrong word for the attitude we should take toward those events." Since the principle of charity applies to our dispositions as well as our actual past utterances, this quick and easy solution to the coherence problem rests on a mistaken application of semantic principles.

The linguistic view of probability is rather slippery and difficult to work with, since its ground rules are not entirely clear. (For example, should folk utterances of "There's only one world" be charitably interpreted as true? Disbelief in the many-worlds ontology may become absurdly difficult if our linguistic practices are interpreted too "charitably.") Given this slippery nature, I am unsure of the linguistic view's implications for free will. It may be that the arguments of this paper could be evaded if the linguistic view were true.

But because I am confident the linguistic view is false, I will set it aside at this point without further comment. Readers left unpersuaded by Bacciagaluppi and Ismael's argument should simply consider the falsehood of the linguistic view to be one of my argument's premises. I cannot rule out the compatibility of many-worlds and free will should the linguistic view of probability turn out, despite appearances, to be correct.

## 2.3 Probabilistic causation and Sebens's argument

Control over one's actions is normally thought to be a prerequisite for free will, and causal influence is a necessary condition for any reasonable notion of control. It therefore behooves us to ask how we might make sense of causation within a many-worlds framework.

One sort of causation within many-worlds is simple to understand, namely what one might call global causation, or the causal determination of the universe's quantum state at one time by the state in the past. At the level of the universe's state, many-worlds is a deterministic interpretation, so the present state is straightforwardly causally necessitated by the past state. Similarly, when decoherence occurs, the whole class of branches splitting off from the state of your present branch is uniquely causally determined by the past within your branch.

More complicated is the question of causation within a branch. When is an event within a branch, such as a single person's decision, caused by another event or state of affairs within that branch? A natural approach for Everettians would be to treat this as a question about probabilistic causation, adopting the SU, OD or PMU account of probability and positing that $C$ causes $E$ when $C$ raises the probability of $E$ and $E$ then occurs. Such an account will require the Everettian to make sense of probabilistic retrodiction, explaining past events on the basis of their probabilities. But an example due to Charles Sebens threatens the prospects for such retrodiction.

A viable account of Everettian causation ought to be able to ground probabilistic explanations of past events. For example, when Davisson and Germer first observed interference in an electron double-slit experiment, we ought to be able to say that the interference pattern was observed because it was highly likely to result from the state they prepared.

Unfortunately Sebens's example shows that Everettian probability cannot be applied to past events, at least not with full generality. He poses a case in which you discover your grandma's old safe deposit box, which was locked away before you were born. Grandma's diary relates that she performed a double-slit experiment and hid the results inside the deposit box. Before looking in the box, can you predict that you are likely to see the familiar interference pattern in grandma's results?

Not if you are an Everettian. For Everettian accounts of probability to apply to an event, it must be the case that you have undergone branching prior to that event–that is, there must be an earlier time such that (a past stage of) you were initially present in a branch, that then branched into further branches, some including the event in question and some not including it. This is a common feature of decision-theoretic proofs of the Born probability law in Everett. Laying out his canonical proof, Wallace sets the goal

> to prove, rigorously and from general principles of rationality, that a rational agent, believing that (Everett-interpreted) quantum mechanics correctly gives the structure and dynamics of the world and that the quantum state of his branch is $|\psi\rangle$, will act for all intents and purposes as if he ascribed probabilities in accordance with the Born Rule, as applies to $|\psi\rangle$. (Wallace, 2012, 159)

The problem is that in Sebens's grandmother example, you (going through the safe deposit box) are not in the position of believing that the state of your branch is the state measured by your grandma. Nor were you ever the inhabitant of a branch with that state. So although Wallace's theorem has the power to establish that your grandma would be rational to assign probabilities to her measurements in accord with the Born rule, the theorem cannot establish that those probabilities are rational for you. Since all of the familiar decision-theoretic and epistemic aspects of probability are supposed to follow from this result about rationality, you are not in a position to expect that your grandma's experiment found the typical interference pattern.

Sebens's argument is interesting for many reasons, but for present purposes its most important upshot is that probabilistic retrodiction is possible in Everettian quantum mechanics only in special cases. And if it is impossible to retrodict from the theory (interpreted Everett-style) that your grandma's two-slit experiment probably produced the typical interference, it seems highly unnatural to say that the theory admits a picture of probabilistic causal explanation that can explain her results in terms of the state she prepared. Although the overall state of the universe can be explained in terms of the past state, it is far from clear that events within an individual branch can be explained in terms of that branch's past.

This is by no means the last word on the subject of probabilistic causal explanation in Everett, but the difficulty of finding room for such explanations presents one clear obstacle in the way of accommodating compatibilist accounts of free will. We'll discuss this obstacle a bit further in the next section, but already one problem should be fairly clear. Compatibilist accounts tend to rest on claims about the ways in which our mental states bring about our actions. Obviously these are claims about the causation of events (within a branch) by other events. If it's impossible for causal explanation to get off the ground in many-worlds at all, it's very hard to see how compatibilist notions of free will could gain any purchase.

## 3 Freedom in many-worlds

With the relevant features of the many-worlds interpretation in mind, we are ready to examine Egan's reasons for doubting whether free will can exist in many-worlds. As noted above, Egan's picture of free will sounds rather compatibilist. Freedom, he claims, consists in choosing in accord with your nature. This sounds like a criterion that could easily be satisfied by agents in a deterministic world.

Indeed, Egan's one-sentence description of his (or at least, his narrator's) view sounds quite a bit like one of the most popular versions of compatibilism: the *deep self* view. Susan Wolf describes it thus:

> [T]he key to responsibility lies in the fact that responsible agents are those for whom it is not just the case that their actions are within the control of their wills, but also the case that their wills are within the control of their *selves* in some deeper sense. (Wolf, 2003, 375)

(For present purposes, I'll assume an agent is free if she is the sort of being who can be responsible for her actions.) The deep self view is appealing because, unlike some earlier, less-sophisticated compatibilist theories, it can convincingly explain why we sometimes excuse an agent from responsibility even for intentional acts. An agent with Tourette's syndrome may compulsively utter obscenities even though her deepest desire is to resist the compulsion. In some sense these utterances are intentional, but they don't reflect who the agent is in the way that considered decisions or long-term plans might.

The earliest antecedents of the deep-self view focused on higher-order desires. Frankfurt (1971), for example, argues that an agent's action is free if it accords with that agent's highest-order desires. An addict may want to get high, but may also desire not to want to get high–if so, the act of getting high is not freely chosen. On the other hand, if the addict's highest-order relevant desire were that she desire to do drugs, this same act would count as freely chosen. The identification of the deep self with an agent's highest-order desires has since been questioned (Watson, 1975), but for present purposes something like Frankfurt's model can serve as a stand-in for a more sophisticated version of the deep self view.

That said, one particular complication will be relevant for our purposes. Consider an agent who acts completely at random, but (by sheer luck) ends up performing an action endorsed by her deep self. Surely this wanton action should not count as freely chosen. This example underscores one of the features of the deep self view described by Wolf above: to be free, the will must be *controlled* by the deep self. That is to say, our free decisions must in some sense be ultimately produced by the deep self–caused by the deep self in the right way.

Quite a lot could be said here about what sort of control is necessary for freedom. One thing to note is that causation will surely be required, and hence the argument of Sec. 2.3, proceeding from Sebens's example, will undermine the possibility of free will unless it can be answered.

It will also be important to note that control is not simply a matter of causal dependence alone. This fact is an important consequence of a classic example in the theory of action:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never *chose* to loosen his hold, nor did he do it intentionally. (Davidson, 1973, 154)

This sort of scenario is called a *wayward causal chain.* Because the climber's mental state causes his behavior in this indirect way, the behavior does not count as under his control even though it fulfills his preferences. We say that the action is not caused by the mental state in the right way. Similarly, if your deep self were to cause your behavior in a wayward manner, the behavior should not count as freely chosen even if it accords with the deep self's motivations.

The significance of this point is simply that our behavior must be caused by our deep selves *in the right way* in order to be free. One way for causation to go the wrong way is for it to proceed via a wayward chain. But in branching realities like the one described by many-worlds quantum mechanics, it is possible for causation to go wrong in another way.

## 3.1 Fission choice examples

Suppose an agent, Boutros, faces a moral decision. In an ordinary, classically deterministic world, Boutros's decision will count as free on the deep self picture if it is ultimately produced from or originates in his deep self. Setting aside the question of what should count as the right sort of production or control, we may say that Boutros's act is free if his deep self endorses it and produces it.

What should we say about a decision made in a branching universe like the one described by the many-worlds interpretation? The branching in many-worlds is closely analogous to Parfit-type fission examples, so it will be useful to consider such an example as an intermediary case. Suppose the following is true:

**Classical Basic Fission.** When faced with the moral decision, Boutros will deterministically fission into two people: Good Boutros, who does the right thing, and Bad Boutros, who does the wrong thing. Both people will be equally causally connected to the original Boutros. There is no morally relevant further fact.

In this case two actions are produced.[1] Is either action (or both) chosen freely?

The answer is no, I would argue. This strikes me as a case in which Boutros's actions lie outside the control of his deep self. It may still be that one of the resulting people's actions accord with his deep self, in the sense that his deep self would endorse those actions.

[1] We may imagine that the two duplicates resulting from the fission are instantly placed into duplicate, identical situations. For example, if Boutros faces the decision of whether to feed a starving cat, he is split into two people, one who feeds a starving cat and one who does not (perhaps the cat is simultaneously split into two cats, and so on).

For example, if Boutros is fundamentally a good person then Good Boutros will behave in accord with his deep self. But neither Good nor Bad Boutros's action is (uniquely) produced by his deep self. (It might be more accurate to say that Boutros has the sort of deep self that tends to simultaneously produce good and bad actions[2]–a self that fits in some ways the popular conception of a "split personality.") If I became convinced that my all of own decisions result in this sort of splitting, I would conclude that I have no control whatsoever over the things I do.

Let me go beyond bare intuitions and present an argument that neither post-fission person acts freely. I think the Classical Basic Fission example is analogous to an example constructed by Mele (1998, 1999). In Mele's example, a character named John is trying to resist the temptation to show up late to a meeting. (To do so would be wrong.) We are supposed to imagine a truly indeterministic choice, so that everything (including the laws) up until the moment of John's choice is compatible with both his arriving punctually and his showing up late.

Suppose John succumbs to temptation, and think of his counterpart John$_2$, who is the same as John up until this decision but resists temptation and shows up on time. The difference between John and John$_2$ would seem to be a matter of luck rather than free choice. Although there is nothing external to John that forced him to show up late, there is also nothing *internal* to John that explains why he did not end up like John$_2$. Although Mele does not discuss his example in these terms, one might think that what is missing is the power of John's deep self to produce his actions (or perhaps, to produce them in the right way). Without that power, John is not free.

The Classical Basic Fission example seems relevantly similar to the John case. The difference between Good Boutros and Bad Boutros is, if anything, more obviously a matter of pure luck than the difference between John$_1$ and John$_2$. Instead of figuratively sharing the same past with a possible counterpart, Good Boutros literally shares the same past with an actual Bad Boutros. There is nothing to explain why Good Boutros does not end up like

[2] Or more precisely, simultaneously decides rightly and wrongly with respect to the same option. For having the sort of deep self that reliably produces an action it does not endorse whenever it produces an action it endorses is not strictly incompatible with freedom, as illustrated by an example of Joel Ballivian: An agent who deeply desires ice cream, but also reflexively slaps herself in the face whenever she eats ice cream, still exhibits free will in choosing to eat the ice cream. What is distinctive about the quantum case is that an agent who desires ice cream would simultaneously choose to eat ice cream and not to eat it.

Bad Boutros, except the brute fact of their different decisions.

I have suggested that in the Classical Basic Fission case, Boutros's actions (or the actions of the people resulting from fission) are not under the control of his deep self. One may object that this is not obvious. Consider the case in which Boutros is a good-natured person whose deep self endorses doing the right thing. Good Boutros is causally produced by the original Boutros, and this was certain to happen given his situation prior to the fission. So it is in a sense correct to say that that initial situation (including Boutros's deep self) reliably produces a person (Good Boutros) who acts rightly. Couldn't Good Boutros's action then be under the control of his deep self? The existence of Bad Boutros, who acts wrongly, might be seen as an irrelevant side issue. As long as Boutros's initial state reliably produces *someone* whose actions accord with his deep self, shouldn't that be sufficient for the right sort of control?

In my view, it is clearly not sufficient. In order for your deep self to have control over your will, it is necessary that you are not the sort of person who will reliably choose to act in a way that doesn't accord with your deep self. But in the Classical Basic Fission case, Boutros's initial state will infallibly lead to a decision, on the part of one of his successors, that his deep self does not endorse. The fact that Boutros's "good" successor will act in accord with his deep self is not by itself sufficient for control.

The right way to understand this example, it seems to me, is that it illustrates a further sense in which the deep self's causal influence over the will must be causation of "the right kind." Causal influence by itself is not enough for genuine control. For one thing, the causal influence must not proceed via wayward causal chains. In addition, as Classical Basic Fission illustrates, the influence must not issue in multiple "contradictory" decisions.

At this point I hope I've convinced you that a deep self compatibilist should not ascribe free will to the agent in Classical Basic Fission. If you aren't convinced, either by the analogy to Mele's John argument or by the intuitions I've marshaled, the remainder of the paper is not likely to convince you either! But assuming you share my judgment about this case, let's take a look at some cases that resemble the many-worlds interpretation more closely.

## 3.2 Decisions in many-worlds

Human decision-making in many-worlds isn't quite directly analogous to the simplified case of Classical Basic Fission. But the complications that enter in many-worlds are not, I think, relevant to our question about free will.

Rather than a single picture, the modern many-worlds interpretation presents us with three main possibilities for how to model decisions, corresponding to the three possible pictures of probability. In each case, we may consider a classical analogue which is identical in every respect relevant to the deep self view of free will. Consider first the objective determinism (OD) view, according to which Everettian probability is a measure of how much we should care about what happens in which branches of the universe following a split. To this picture there corresponds the following elaboration of the Classical Basic Fission example:

**Classical Objective Determinism.** When faced with the moral decision, Boutros will deterministically fission into two people: Good Boutros, who does the right thing, and Bad Boutros, who does the wrong thing. Both people will be equally causally connected to the original Boutros. It is objectively rational to consider the interests of Good Boutros to be 999 times as important as the interests of Bad Boutros.

This is analogous to Greaves's claim that "the rational Everettian cares about her future successors in proportion to their relative amplitude-squared measures." (Greaves, 2004, 430) The disanalogies, namely that the fission in Classical OD is explained in classical terms and the "probability" does not arise from the amplitude of the wavefunction, seem clearly irrelevant to the question of who we ought to consider a free and responsible agent.

I think it is also clear that Classical OD is no different from Classical Basic Fission when it comes to this question of freedom. In Classical OD, just as in Basic Fission, one of your post-fission successors may act in accord with your deep self, but the production of the other successor indicates that the post-fission actions are not under the deep self's control. The good successor now matters more than the bad successor. But this does not make it any easier to ascribe the good successor's action (and not the bad successor's action) to your deep self.

If this were an indeterministic event without fission, and Boutros performed the good action and not the bad action, and the *objective chance* of Boutros performing the good action were 99.9 percent, that would be quite a different situation, in at least two ways. First, it could be unambiguously true that Boutros's deep self was the probabilistic cause of his willing the good action–the explanation for why the good action was so probable. This introduces the possibility that the action was under his deep self's control. Second, since Bad Boutros would not exist in that case, it would be possible for Boutros's deep self to cause the good action (or his willing of it) in the right way, without also simultaneously bringing about an inconsistent act of will by Bad Boutros.

But there is no analogous way to describe Classical OD as a case of probabilistic causation, because strictly speaking there is no objective probability in Classical OD–only a measure of how much Good and Bad Boutros's interests matter. And Bad Boutros does, of course, exist in the Classical OD example, ensuring that Boutros's deep self can't properly be said to control his will. I conclude that the Classical OD scenario is incompatible with free will on the deep self view. Since decision-making in many-worlds is analogous to Classical OD in every relevant way if the OD approach to probability is correct, it must be that on the OD approach, the many-worlds interpretation is incompatible with deep-self free will.

The subjective uncertainty (SU) view may appear more promising for free will, but this appearance is illusory. The argument that SU is relevantly analogous to Basic Fission will be more complicated, but in the end I think the analogy holds up and the same problem arises in a slightly different form. Consider the following case:

**Classical Subjective Uncertainty.** When faced with the moral decision, Boutros will deterministically fission into two people: Good Boutros, who does the right thing, and Bad Boutros, who does the wrong thing. More precisely: Boutros's current person-stage is a temporal part of two continuant people, Good Boutros and Bad Boutros. Prior to fission, it is rational for him to believe with degree 0.999 that he is Good Boutros.

This case is similar in every relevant way to quantum SU; the only differences lie in the details of the physical mechanism of fission.

Is Classical SU more compatible with deep-self free will than Classical OD? It may seem so at first. In Classical SU, Good Boutros exists prior to fission (along with Bad Boutros, who shares the same person-stage). So it is in a sense pre-determined that Good Boutros will do the right thing and hence act in accord with his deep self. There is als a sense in which Good Boutros's action should be regarded as more probable than Bad Boutros's action, which points to the possibility of probabilistic causation.

But recall why we deemed Boutros's situation incompatible with free will in the Basic Fission case. The problem was that Good Boutros and Bad Boutros share the same deep self. The fact that the same deep self reliably produces both Good Boutros's right action and Bad Boutros's wrong action is sufficient to establish that neither agent is free. But in Classical SU, we have the exact same reasons for ascribing the same deep self to Good and Bad Boutros. Just like in the Basic Fission case, they share all person-stages up until the moment of decision. There is no quality of character or will that Good Boutros possesses and Bad Boutros lacks (prior to the decision), and yet it is physically predetermined that their actions will diverge.

This means there is no relevant difference between Classical SU and Basic Fission. In Classical SU, we do have the further fact that the initial person-stage (Boutros) ought to believe he is a person identical with Good Boutros and not Bad Boutros. But this doesn't really help matters. For one thing, up until the moment of fission, it isn't clear what could make this belief true or false. The fact that it is (according to SU) a justified belief has no apparent significance for the free will question. If there were some further fact that determined the truth (or objectively probable truth) of this belief, we could then tell a story about how Boutros's deep self is a cause of Good Boutros's action but not Bad Boutros's action. In the absence of such a further fact, though, there is nothing more to say about this case beyond affirming our earlier conclusions about Basic Fission. Because Bad Boutros's action issues deterministically from the same deep self as Good Boutros's action, neither agent is free if we assume the deep-self picture.

We are left with the Post-Measurement Uncertainty (PMU) picture of probability. Recall that on this picture, probability enters only after the worlds split. Prior to the split, there is no sense to be made of the probability that you occupy one branch or another–the branches for distinct outcomes don't exist at that point. In this sense, the PMU picture is the closest to the Basic Fission example. The only significant difference is the introduction of post-branching probabilities, which may be incorporated into the Basic Fission example as follows:

**Classical Post-Measurement Uncertainty.** When faced with the moral decision, Boutros will deterministically fission into two people: Good Boutros, who does the right thing,

> and Bad Boutros, who does the wrong thing. Both people will be equally causally connected to the original Boutros. Immediately following the fission, it will be objectively rational to believe with degree 0.999 that both people are Good Boutros.

This may sound somewhat incoherent. Bad Boutros should be very confident that he is Good Boutros? That is, Bad Boutros should believe he will do the right thing, even though he will do the wrong thing? This seems not to fit with our ordinary conception of agency and decision-making.

This example is easier to understand from the perspective of a bystander. In quantum branching, any bystander will equally be rationally required to apportion their credence in accord with the quantum probability rule. So for the Classical PMU case to be analogous to quantum PMU, it must be rational for a bystander to believe with degree 0.999 that Bad Boutros is Good Boutros (in the moment immediately following the fission). So following Boutros's decision, a bystander who has just met one of the post-fission copies of Boutros, and has not verified whether this copy is Good or Bad Boutros, should believe that this copy is Good Boutros. This is the easiest way to understand Classical PMU.

Classical PMU is analogous to Classical SU, but only after the moment of fission. Thus, if my argument against free will in Classical SU was convincing, you should be equally convinced that Classical PMU is incompatible with free will.

Before the moment of fission, Classical PMU is entirely analogous to Classical Basic Fission. In particular, there is no sense in which it is probable, prior to fission, that Boutros will do the right thing. As in Basic Fission, it is simply the case that (with probability one) Boutros will fission into Good Boutros and Bad Boutros, and that Good Boutros will act rightly and Bad Boutros will act wrong. It is also true that following the fission, it will be rational to believe that Good Boutros is Good Boutros, and also rational to believe that Bad Boutros is Good Boutros. But by themselves, these facts about what it's rational to believe imply nothing whatsoever about whether either Boutros copy's action was produced by Boutros's deep self.

All of the facts about Boutros's deep self, and the effects it has on his successors Good and Bad Boutros, are exactly the same as in Basic Fission. So we have no more grounds to judge that Boutros acts freely in Classical PMU than we do in Basic Fission. The PMU approach is no friendlier to free will than any of the other versions of the many-worlds interpretation.

# 4 Conclusions

I'll close with two observations.

To state what should be obvious, the argument of Secs. 3.1 and 3.2 is no reason to disbelieve the many-worlds interpretation (although it may be a reason to hope the interpretation is false). That said (and this is the first of the two observations I promised), Sebens's example from Sec. 2.3 does point to a more general outstanding puzzle for the interpretation. For this example raises a very broad concern about probabilistic causation in this interpretation of quantum mechanics.

The three pictures of probability provided by the Objective Determinism, Subjective Uncertainty and Post-Measurement Uncertainty views reproduce many of the most important properties that we normally lump under our concept of objective chance. If the OD and SU views succeed, for example, it is clear that rational agents should behave in accord with the probability rules of ordinary quantum mechanics when it comes to decision-making. And on all these views, quantum probability behaves like objective chance for purposes of confirmation theory.

But the defenders of these views have not shown that any of these three concepts of probability can undergird anything like our ordinary notion of probabilistic causation. As Sebens's example reveals, although these three approaches can ground predictions about why we should expect and plan for high-probability future events, none of these approaches can explain why high-probability events have occurred in the past. (They can explain why it was previously rational to expect these events to occur, but that isn't the same as explaining why they occurred, or even retrodicting after the fact that they were likely to occur.)

If this worry about probabilistic causation were solved, and an account of probability advanced that could handle Sebens's example, this would not show that there isn't a special problem for free will in many-worlds. The problem would remain that branching causation seems not to be the *right sort* of causation to undergird the sort of control we need to have over our actions in order for them to be free. But solving the problem about probabilistic causation would at least help to establish the scientific viability of the interpretation. It would mean, that is, that none of the points in this paper count against the truth of many-worlds.

Observation number two: There may be other approaches to compatibilism that are more hospitable to many-worlds. I would caution against undue optimism, however. The deep self account is perhaps the most widely-held contemporary compatibilist picture, and arguably the only one to provide anything like a satisfactory answer to incompatibilist objections. Making compatibilism work is a very difficult project, and to throw out the most promising theory may mean giving up the game.

That said, it may be that the deep self view can be modified to accommodate many-worlds if we alter our notion of what we freely control (that is, which facts and events we are responsible for). Perhaps we are responsible, not for the individual actions our future selves take within individual branches, but for the overall pattern of our successors' future actions and their quantum weights (which we ordinarily think of as probabilities but may or may not really deserve that name). The deep self does influence this overall pattern: a good-natured personality will increase the weight of future branches where the agent acts rightly. This is not the ordinary sort of freedom we take ourselves to have, which surely requires having control over our individual actions. But it may be a form of freedom that is worth wanting, if many-worlds provides the correct picture of our universe and its structure. And if that is the sort of universe we live in, this may be the best sort of freedom we can hope to possess.

# References

Bacciagaluppi, Guido and Jenann Ismael (2015), "Review of The Emergent MultiverseDavid Wallace, The Emergent Multiverse: Quantum Theory According to the Everett Interpretation. Oxford: Oxford University Press , Xvi+530 Pp., $75.00," *Philosophy of Science* 82:129–148.

Davidson, Donald (1973), "Freedom to Act," in Honderich, Ted (ed.), *Essays on Freedom of Action*, Routledge.

Deutsch, David (1999), "Quantum theory of probability and decisions," *Proceedings of the Royal Society of London* A455:3129–3137.

Egan, Greg (2002), "Singleton," *Interzone* 176.
URL http://gregegan.customer.netspace.net.au/MISC/SINGLETON/Singleton.html

Frankfurt, Harry G. (1971), "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68:5–20.

Greaves, Hilary (2004), "Understanding Deutsch's Probability in a Deterministic Multiverse," *Studies in History and Philosophy of Modern Physics* 35:423–456.

Lewis, David (1976), "Survival and Identity," in Amelie Oksenberg Rorty (ed.), *The Identities of Persons*, University of California Press, 17–40.

Mele, Alfred R. (1998), "Review of Robert Kane," *The Journal of Philosophy* 95:581–584.

Mele, Alfred R. (1999), "Ultimate Responsibility and Dumb Luck," *Social Philosophy and Policy* 16:274–293.

Parfit, Derek (1971), "Personal Identity," *Philosophical Review* 80:3–27.

Saunders, Simon and David Wallace (2008), "Branching and Uncertainty," *British Journal for the Philosophy of Science* 59:293–305.

Sebens, Charles T. and Sean M. Carroll (forthcoming), "Self-Locating Uncertainty and the Origin of Probability in Everettian Quantum Mechanics," *British Journal for the Philosophy of Science* .

Vaidman, Lev (1998), "On Schizophrenic Experiences of the Neutron or Why We Should Believe in the Many-Worlds Interpretation of Quantum Theory," *International Studies in Philosophy of Science* 12:245–261.

Wallace, David (2003), "Everettian Rationality," *Studies in History and Philosophy of Modern Physics* 34:87–105.

Wallace, David (2012), *The Emergent Multiverse*, Oxford: Oxford UP.

Watson, Gary (1975), "Free Agency," *Journal of Philosophy* 72:205–20.

Wiggins, David (1967), *Identity and Spatio-Temporal Continuity*, Oxford, Blackwell.

Wolf, Susan (2003), "Sanity and the Metaphysics of Responsibility," in Watson, Gary (ed.), *Free Will*, Oxford: Oxford UP, 372–387.