Epistemology Quantized: circumstances in which we should come to believe in the Everett interpretation

David Wallace*

July 2005

Abstract

I consider exactly what is involved in a solution to the probability problem of the Everett interpretation, in the light of recent work on applying considerations from decision theory to that problem. I suggest an overall framework for understanding probability in a physical theory, and conclude that this framework, when applied to the Everett interpretation, yields the result that that interpretation satisfactorily solves the measurement problem.

1 Introduction

Recent years have seen substantial progress on both of the main problems traditionally associated with the Everett interpretation.

The first of these, the 'preferred basis problem' is not my concern here; suffice it to say that I believe considerations from decoherence theory, together with the right philosophical analysis of higher-order ontology, appears sufficient to resolve it.[1] My concern is with the second, the 'probability problem', where again a combination of technical and conceptual results have transformed a problem which appeared intractable.

The probability problem can usefully be divided into two parts:

The incoherence problem: In a deterministic theory where in theory we might have perfect knowledge, how can it even make sense to assign probabilities to outcomes?

The quantitative problem: Even if it does make sense to assign probabilities to outcomes, why should they be the probabilities given by the Born rule?

*Magdalen College, Oxford

[1]For a more detailed analysis see Wallace (2003a).

Until fairly recently, both problems seemed intractable: any attempt to resolve either, at least without modifying the basic structure of quantum mechanics seemed doomed to failure. But recent work by Deutsch (1999), Saunders (1998), Vaidman (1998), Greaves (2004), myself (Wallace 2003b,Wallace 2003c,Wallace 2005a,Wallace 2005b) and others has made use of considerations from decision theory, personal identity, philosophy of probability, and philosophy of language to provide both conceptual frameworks for thinking about probability in the Everettian context, and — perhaps more surprisingly — concrete mathematical results which purport to be (Everett-interpretation-specific) derivations of the Born rule.

Intractability has a certain simplicity. By contrast, our current state of knowledge about the probability problem has become quite complex and controversial; it is unclear exactly how to frame the problem which we are trying to solve, and correspondingly unclear what would count as a solution.

This paper is an attempt to untangle the situation: in it I have tried to lay down exactly what I believe is involved in a solution to the measurement problem, and to show what needs to be done for the Everett interpretation to be understood to provide such a solution.

Section 2 is concerned with probabilistic theories in general and not with quantum mechanics (let alone the Everett interpretation) in particular; in this section I advocate what I call 'cautious functionalism' as the correct attitude to take to probability in physical theories.

In the remainder of the paper, I apply this framework to quantum mechanics and the Everett interpretation. Section 3 lays out my preferred approach to the Everett interpretation: in sections 3.1–3.2 I set out a framework for what is required of a solution to the measurement problem and how that framework applies to the Everett interpretation (in particular, how it requires that we understand quantum branching events as genuinely uncertain in some sense). Sections 3.3–3.5 analyse how uncertainty can be understood to exist in a branching universe, and section 3.6 considers how (given an understanding of uncertainty) quantitative probability can also be understood in a branching universe — understood better, in fact, than in a non-branching universe!

In section 4 I consider an alternative approach: the 'fission program', in which we make no essential use of the concept of uncertainty or subjective probability in considering branching. Although I conclude that the fission program is ultimately not a viable approach to the Everett interpretation, it does lead to some important insights which become relevant in section 5, in which I consider whether the decision-theoretic principles which are part and parcel of the 'uncertainty' concept are really justified in a branching universe. Section 6 is the conclusion.

2 What is probability?

2.1 Objective probability and the principal principle

It is fairly widely accepted that there are (at least) two distinct notions of 'probability'.

Firstly, there is subjective probability, or 'credence': that is, probability taken as a measure of an agent's degree of belief in a hypothesis. Here, probability is taken as a numerical quantification of the notions of 'likelihood' and 'uncertainty', which arguably we already understand in qualitative terms.

Foundationally speaking, credence is in fairly good shape. If asked to justify its use, our first response is to appeal to so-called 'Dutch Book' arguments to show that any other method of quantifying our beliefs forces us to lose money. On careful analysis these are not really convincing, but they have far more convincing relatives in *decision theory*. Decision theory (in the forms developed by Savage (1972) and Jeffrey (1983)) begins with a purely qualitative notion of an agent's preferences over various courses of action, and allows us to prove a *representation theorem* to the effect that that agent must quantify his uncertainty in terms of probability or violate some intuitively reasonable principle of rationality. (For a brief but careful defence of this approach to credence, see the first chapter of Kaplan (1996); for more details, see Joyce (1999)).

But if credence is well defined, it is nonetheless too thin a notion to play the role of *objective* probability (OP), the robust, observer-independent property which we use in much of science and in particular in quantum mechanics. We have, I believe, no truly satisfactory analysis of what sort of entity or property this 'objective probability' really is.

We do, however, have a good theory of how OP fits into our general conceptual scheme: it does so via Lewis's *Principal Principle* (Lewis 1980). Recall that the Principal Principle states, roughly, that if I know the objective probability of an event E to be p, then I am rationally compelled to set my personal credence in E to be p. More precisely, it states that if X_p is the proposition that $OP(E) = p$, and A is any 'admissible'[2] proposition compatible with X, then

$$Cr(E|A\&X_p) = p. \tag{1}$$

(Lewis uses 'chance' as his term for OP, and regards it as necessarily involving indeterminism: classical statistical mechanics, and Bohmian mechanics, aren't chancy in his sense. I shall use the term more broadly: the probabilities of statistical mechanics seem as robust and as observer-independent (and as mysterious!) as those of stochastic theories, and (provided that 'inadmissible propositions' is extended to cover microphysical knowledge such as the actual microstate of the system or the actual location of the Bohmian corpuscle) the Principal Principle applies just as well for them. There is no particular etymological reason to restrict 'chance' to chances in Lewis's sense, but to avoid

[2] Inadmissible propositions are not formally defined by Lewis, but essentially the qualifier is there to rule out propositions which are directly about the future — e. g. , prophecies or the testimony of time travellers.

confusion I use the more neutral, albeit more cumbersome, 'objective probability' to cover the more general class of physics-defined probabilities.)

Other than that it satisfies the Principal Principle, what do we actually know about OP? Answer: essentially nothing. *Mathematically*, it enters through either a measure on the set of initial conditions or a stochastic differential equation — that is, in effect, through a measure on either the initial conditions or on the dynamically possible histories. But the interpretation of that measure remains obscure. There are proposals — such as frequentism, or Lewis's Best-Systems analysis — that try to define that measure in terms of facts about the actual world, but it is at best extremely controversial both whether these proposals allow us to define an appropriate measure at all, and whether it could be shown to constrain a rational agent's credences.

This being the case, we may as well take the Principal Principle as offering a *functional definition* of OP. That is, if some physical theory T enables us to define some magnitude C for events, then C is OP just if anyone believing T is compelled to constrain his credences to equal C. More formally, C is OP iff for any event E, if T together with (admissible) background information B entails that $C(E) = p$, then

$$Cr(E|B\&T) = p. \tag{2}$$

This 'functional definition' allows us to sidestep — temporarily — the question of what chance is when we consider its role in science. For suppose we do have a theory T which allows us to define some magnitude C, and suppose PP_C is the proposition that C satisfies the functional definition (we might loosely say: satisfies the Principal Principle — hence the notation). It follows that if we accept both T and PP_C, we should set our credence in an event E equal to $C(E)$. If $C(E)$ is high and our prior credence in E is much less high, we should regard $(T\&PP_C)$ as explanatory of E, and thus regard E as reason to accept T and PP_C.

This argument can be rephrased in Bayesian terms:

$$Cr(T\&PP_C|E) = \frac{Cr(E|T\&PP_C)Cr(T\&PP_C)}{Cr(E)}; \tag{3}$$

hence

$$\frac{Cr(T\&PP_C|E)}{Cr(T\&PP_C)} = \frac{C(E)}{Cr(E)}. \tag{4}$$

Since to accept both T and PP_C is to accept the existence of objective probability, it follows that we can gain evidence — even very powerful evidence — for objective probability functionally defined, without ever knowing what sort of thing that 'objective probability' really is.

2.2 Three ways of satisfying the functional definition

Nonetheless, we'd still like to *know* what it is. What *is* the "magnitude C" defined by the theory T? There are essentially three options. The first, which might be called *functionalism*,[3] asserts that probability is some physically-definable property (or set of physically-definable properties) which can be defined independently of the Principal Principle but which can be shown to satisfy the functional definition which the Principle provides.

According to functionalism, PP_C must somehow be a logical consequence of T (more precisely: of T together with general principles of rationality and possibly other background assumptions). For instance, a frequentist (put crudely) identifies probability with long-run relative frequency; he is correct iff it can be proved that a rational agent knowing the long-run relative frequency of a particular outcome of a repeated experiment would set his credence in getting that particular outcome on a single run equal to the long-run relative frequency.

The problem with frequentist versions of functionalism is that we have very little idea how to prove that long-run frequency can be proved to satisfy the functional definition of probability; the problem with functionalism more generally is that we have very little idea how to prove that *anything* can be proved to satisfy that definition. The problem is not so much that we don't know how to define something with the formal properties of probability (relative frequencies do okay here; Lewis's 'best-systems analysis' (Lewis 1994) does better); rather, it is that we do not have any really plausible account of why that something should place any constraints on my credences. Why should I, betting in the here and now on whether *this* atom is going to decay, care at all about how many similar atoms in remote regions of the Universe have decayed?

The apparent impossibility of finding a naturalistic candidate for probability provides much of the attraction for the second option, *primitivism.* Primitivists accept the functional definition as a basic law of nature: not something to be deduced as holding for an independently characterisable property C, but something which is postulated to be true of C and which defines C via its role in the law.

The strategy is not unfamiliar. Take *charge*, for example: it is highly plausible to suppose that the property of having charge q cannot be defined or made sense of other than via the role of charge in the laws of electromagnetism (see Lewis (1970) for a full working-out of this strategy[4]). Furthermore, it seems to fit well with the mathematical structure of our existing probabilistic theories: as was alluded to above, both in stochastic theories like GRW and in deterministic theories with an unknown microstate, objective probability eventually enters as a measure on the space of physically possible histories, which has no role in the theory except to be probability: that is, to satisfy the functional definition.

Nonetheless, primitivism is a desperate strategy. Do we really want to take a *rationality principle* as a basic postulate of nature, on a par with the dynamical laws of spacetime and field theories? Are we prepared to accept that it is logically possible that every physical property of the universe could remain the

[3]Functionalism is close to the position espoused by Lewis (1986, xiv-xvi), who requires that probabilities be shown to be 'Humean properties'.

[4]See see Shoemaker (1980) and (Mellor 1991) for further defences of this 'functional' definition of properties; note also that Lewis seems later (Lewis 1983) to have moved to a different position.

same in some alternate possible world, and yet that what is rational could change?

The third strategy — *eliminativism* — is in turn motivated by the grave conceptual problems of the other two strategies. It is the doctrine that objective probability does not exist: that it is unsurprising that we cannot work out what fits the functional definition of probability, for nothing does.

If primitivism is desperate, eliminativism is all but unacceptable. The ubiquity of the concept of objective probability throughout science, and indeed in ordinary life (think of the roulette wheel and the fair coin) makes it intolerable not to accept that *something* fits the functional definition given by the Principal Principle. (Or so it seems to me; but others disagree. How they can maintain with a straight face that *the half-life of uranium* is not an objective property of the world is beyond me, but I shall not attempt to defend the point further here.)

2.3 Cautious functionalism

> Philosophers' Syndrome: mistaking a failure of imagination for an insight into necessity (Dennett 1991, p. 410)

Faced with this dispiriting trilemma, what attitude should we take towards probability statements in our physical theories? I suggest a cautious functionalism. Unlike the other two strategies, the only philosophical problem with functionalism is our total inability to think of anything that might fit the functional definition ... but our imagination has failed before.

Cautious functionalism, faced with a theory T that defines some property C which seems to play the role of probability, proceeds as follows. It collects evidence, as above, for the joint hypothesis ($T\&PP_C$), all the while acknowledging that T comes with an attached promissory note: eventually we will need an account both of how C is to be defined independently of the Principal Principle and of how, given this independent characterisation of C, PP_C can be derived. Until the note is cashed, the theory has a certain phenomenological, non-fundamental character, yet for all that it may be highly explanatory. And if the note is never cashed, perhaps eventually we would be wise to become pessimistic and reconsider primitivism or (*just* possibly) eliminativism.[5]

[5]It may seem somewhat strange that PP_C, which according to functionalism must in principle be derivable *a priori*, is nonetheless the sort of thing for which we can collect evidence. But there is nothing strange about having *a posteriori* evidence for *a priori* truths. (How many of us have actually worked through the proof of Fermat's last theorem, rather than trusting the word of others that it is provable?) There is not even anything particularly strange about gaining *a posteriori* evidence for a normative principle. For example, consider the Monty Hall problem: a game-show featuring three doors, with goats behind two of them and a car behind the third; you choose one door, and without revealing what is behind it the host opens one of the other doors to reveal a goat. Two doors remain; if offered the chance to open your original choice or the other one, should you swap? Yes, in fact; but remarkably many people both get the problem wrong and resolutely refuse to believe that they *have* got it wrong. Such a person might watch many reruns of the game show, and conclude that in fact he *should* swap, and yet be at a loss to understand *why* he should swap.

Cautious functionalism is not actually (I hope!) that contentious a position. Primitivism and eliminativism are not views that one would adopt despite having a perfectly satisfactory functionalist candidate for objective probability; they are views that one adopts in response to the belief that nothing could be such a candidate. (For instance, objections to frequentist definitions of probability are made on the grounds that they *don't work* (that is, in my framework, that they don't fit the functional definition) — not that even if they did work they would be unsustainable.) So a primitivist, or an eliminativist, can be seen as a cautious functionalist who has already reached the point of pessimism and given up on any successful functionalist analysis. Either, I hope, would willingly recant their pessimism if shown that a functionalist analysis was after all possible.

2.4 Is the functional definition complete?

Before turning from probability in general to the specific case of quantum mechanics, I must address one possible worry: can we really be sure, in using the Principal Principle to produce our 'functional definition of probability', that we have not left out a crucial feature of objective probability? Maybe there is some additional feature F of probability, such that something satisfying the functional definition but not possessing F would not be objective probability.

As a matter of semantics, this may well be defensible — our ordinary 'probability' talk is inchoate and ambiguous between objective probability and credence, and maybe it does have additional features and complications. But if we are discussing the *scientific conception of objective probability* — that is, the theoretical term which we have introduced to explain experimental situations which seem to need probability — then I am not sure what evidence we could have for F.

To see this, let us temporarily introduce "quasiprobability" as a term for anything satisfying the functional definition given by the Principal Principle and discussed above. It is, I take it, uncontentious that probability (whatever it is) satisfies the definition[6], so probability is a certain sort of quasiprobability — one possessing the additional feature F which was left out of the functional definition. Suppose we have a collection of experimental data which is explained well by some theory T involving 'genuine' probability. That is, T assigns high probability to the relative frequencies which we in fact observe. Suppose also that we have some other theory T' which involves only 'quasiprobability' but which assigns high quasiprobability to the observed relative frequencies — in fact, which assigns quasiprobabilities exactly equal to the 'real' probabilities assigned by T.

In such a situation the extra property F appears quite redundant. The evidential process by which we continue to test T and T' connects (quasi-)probability to our observations entirely through the Principal Principle, which applies to

[6] For an argument to this effect, see Lewis's "questionnaire" Lewis 1980; similar arguments have been advanced by, e. g. , (Mellor 1971).

quasi-probability whether or not it has the property F. And the same will apply when we come to use the theory in practical applications — to go from (quasi-)probabilities to actions, we must go via the Principal Principle.

The only use that I can see for F is that it might be a necessary part of *why* probability satisfies the functional definition. If, for instance, relative frequencies were in fact the only possible candidate for a realiser of the functionalist definition of probability, then something with the formal properties of probability which is not a relative frequency could not be true probability.

But this is only to say that our initial guess that a particular quantity *is* a quasi-probability could be wrong. If it is wrong, if that quantity does not satisfy the functional definition, then it is no quasi-probability at all. And if some quasi-probability *could* be shown to satisfy the functional definition (as, I shall argue, occurs in the Everett interpretation) whilst lacking F, then it would be a demonstration that F was not after all necessary.

If there are requirements for probability over and above the functional definition, I conclude that it is obscure at best what they could be. For the rest of this paper, then, I shall assume that quasi-probability is a redundant concept. Anything which genuinely does satisfy the functional definition is probability.

3 The Everett interpretation and subjective uncertainty

3.1 Interpreting quantum mechanics

I have suggested a general framework for the understanding of theories which incorporate objective probability. But what of quantum mechanics? There the theory seems to speak of objective probability, but there also the theory seems ill-defined without an 'interpretation'.

As an interpretation-neutral approach to this question, I suggest the following: what we currently possess is a *theory fragment*. To be more precise: quantum mechanics can be understood as giving a description of certain (usually microscopic) systems. But the connection between that description and our observations proceeds not via a continuation of the theory but via an *algorithm*: when a state comes to describe a superposition of macroscopically-definite outcomes, reinterpret the mod-squared-amplitudes of each outcome as giving the objective probability of that outcome's obtaining.

Regarded this way, this 'theory fragment' (which we might call the Quantum Algorithm) is deficient in two ways, the second by far worse than the first. Firstly, in admitting objective probabilities without providing a theory of same, it lacks a truly acceptable account of probability. However, as the previous section argues, it shares this defect with all other probabilistic theories, so perhaps calling it a 'defect' at all is unfair.

Secondly, and more pressingly, it is not really a theory capable of describing reality in an observer-independent way. Rather, the 'reinterpretation' of the quantum state which occurs when it becomes macroscopic has no real explanation attached to it beyond its empirical success. That is: the theory fragment gives no principled answer to the question of why the reinterpretation should be made; the only justification is that it seems to make correct predictions.

What is the goal of an 'interpretation' of quantum mechanics? I claim that it is to embed the Quantum Algorithm into a genuine theory, one which does not resort to pragmatic considerations and 'reinterpretation' of its basic ontology. This embedding can be exact (as the Everett interpretation claims to be) or approximate (as in dynamical-collapse theories), and it may add additional structure (as in hidden-variable theories), but in any case it must be sufficiently accurate that it can reproduce the powerful empirical success of the Quantum Algorithm.[7]

So: our interpretation must resolve the ambiguity and ill-definedness inherit in the move from micro to macro. Must it also offer a physical property that satisfies the Principal Principle? It depends on how strict our criteria are for an acceptable physical theory. The previous section, in advocating 'cautious functionalism', argued that *eventually* we must move from treating the Principal Principle as a primitive rule and begin to treat it as a derived result, but no probabilistic theory so far proposed has met that stringent test. For the reasons given in sections 2.1–2.4, I would maintain that *inter alia*, an interpretation could take the Principal Principle as primitive for whatever physical properties it likes, and yet be in as secure a position as any other probabilistic theory which physicists treat seriously.

3.2 The need for subjective uncertainty

What of the Everett interpretation? It aims to embed the Quantum Algorithm (that is, instrumentalist quantum mechanics) into a full theory in the most naive possible way: that is, by extending the formalism which we use for microscopic systems to cover all systems, be they microscopic, macroscopic or cosmological. To do so, it must interpret the macroscopic superpositions that result post-measurement as describing a superposition of different 'worlds': different quasi-classical structures in the wave-function, effectively isolated from one another, with some worlds corresponding to each possible outcome of the experiment.

It is not the task of this paper to consider how this many-worlds description of macroscopic superposition may be justified; I present my own proposed justification in Wallace 2003a. But even if it can be done, there remains a seemingly insurmountable problem: how can the Quantum Algorithm, which involves objective chance, be incorporated into a deterministic theory in which an agent could in principle have perfect information about all of the salient features of

[7] I ignore here the possibility that the Quantum Algorithm itself is misunderstood by interpreting it as speaking of objective probability. This possibility has been defended by (amongst others) Chris Fuchs (Fuchs and Peres 2000; Fuchs 2002), who wishes to regard the quantum state as some sort of credence function rather than something which is in any sense objective. This would lead to a theory of probability of the sort which I earlier called 'eliminativist' and which I criticised for taking science insufficiently seriously; I shall not discuss it further here.

the quantum state?

To elaborate: the Quantum Algorithm assigns objective chances to the possible outcomes of quantum-mechanical experiments, and the defining feature of an objective chance is that rational agents are compelled to set their credences equal to it (if they know it). Thus, the Quantum Algorithm assumes that agents have credences in the different outcomes, which in turn are to be understood as quantitative measures of how certain or uncertain they are about the result of the experiment.

But if the Everett interpretation is true, what is there to be uncertain about? The interpretation makes a deterministic prediction about the post-experiment state: namely, that it consists of many effectively isolated worlds, with different measurement outcomes occurring in different worlds. How can we make sense of being uncertain of the outcome of an experiment in a deterministic theory where we have perfect knowledge?

I have elsewhere called this the problem of *subjective uncertainty*. (The 'subjective' should not be taken too literally: the problem is to understand why uncertainty statements can rationally be made by agents embedded in the Everett universe despite their total knowledge of the relevant facts, but these statements could be 'there might be a sea-battle tomorrow' just as readily as 'I don't know what result I will see'.)

3.3 Saunders' argument for subjective uncertainty

The idea of subjective uncertainty (though not the name) was originally proposed by Saunders (1998), who argues for the SU viewpoint by means of an ingenious intuition pump.[8] His argument proceeds by analogy with "classical splitting", such as that which would result from a Star Trek matter transporter or an operation in which my brain is split in two. It may be summarised as follows: in ordinary, non-branching situations, the fact that I expect to become my future self supervenes on the fact that my future self has the right causal and structural relations to my current self so as to *count* as my future self. What, then, should I expect when I have two or more such future selves? There are only three possibilities:

1. I should expect abnormality: some experience which is unlike normal human experience (for instance, I might expect somehow to become both future selves).
2. I should expect to become one or the other future self.
3. I should expect nothing: that is, oblivion.

[8] I should note, in criticising Saunders, that his thought experiment was intended primarily to argue against the claim that quantum branching is metaphysically incoherent, and only secondarily to defend subjective uncertainty. I have no quarrel with Saunders' primary goal; he has conclusively established that the 'metaphysical incoherence' argument is indefensible, given that analogous situations could perfectly well occur in classical physics, and I discuss the matter no further here.

Of these, (3) seems absurd: the existence of either future self would guarantee my future existence, so how can the existence of *more* such selves be treated as death? (1) is at least coherent — we could imagine some telepathic link between the two selves. However, on any remotely materialist account of the mind this link will have to supervene on some physical interaction between the two copies — an interaction which is not in fact present. This leaves (2) as the only option, and in the absence of some strong criterion as to which copy to regard as "really" me, I will have to treat the question of *which* future self I become as (subjectively) indeterministic.

(In understanding Saunders' argument, it is important to realise that there are no further physical facts to discover about expectations which could decide between (1-3): on the contrary, *ex hypothesi* all the physical facts are known. Rather, we are regarding expectation as a higher-level concept supervenient on the physical facts — closely related to our intuitive idea of the passage of time — and asking how that concept applies to a novel but physically possible situation.)

Of course (argues Saunders) there is nothing particularly important about the fact that the splitting is classical; hence the argument extends *mutatis mutandis* to quantum branching, and implies that agents should treat their own branching as a subjectively indeterministic event.

3.4 Objections to Saunders' argument

In responding to Saunders, Greaves (2004) argues as follows:

> What (to address Saunders' question) should [someone about to be duplicated] *expect* to see? Here I invoke the following premise: whatever she knows she will see, she should expect (with certainty!) to see. So she should (with certainty) expect to see [herself as the first duplicate], and she should (with certainty) expect to see [herself as the second duplicate]. Not that she should expect to see *both*: she should expect to see *each*. (Greaves (2004, p. 19); quotation modified to remove subscripts on pronouns, which will play no role here.)

As Greaves freely admits, Saunders is unwilling to accept this extra possibility, on the grounds that it is just conceptually impossible to expect two incompatible possibilities. She responds by claiming that it is conceptually impossible to feel uncertain about something when one knows all the facts about it.

I think that this impasse can be clarified (if not resolved) by interpreting Greaves as raising the possibility of *concept failure*: a breakdown of our concept of personal identity (Saunders, of course, has implicitly assumed that this concept remains applicable in cases of splitting and seeks, via his possibilities (1)–(3), to ask *how* exactly it is applicable.).

Many of Parfit's examples (in Parfit 1984, pp.199–306) are designed to suggest the possibility of concept failure: Parfit's intention in doing so is to persuade the reader to *give up* on personal identity as something worth caring about and to replace it with a notion of *personal survival* (according to which it is perfectly coherent for me to care about my future successors without having any particular view about whether they are *me* or not). I think that Greaves is best read as sharing this view: according to her version of the Everett interpretation, I should replace any notion of *becoming* a post-splitting version of myself simply with the notion of *caring about* the future versions of myself, and should treat 'I expect experience X' simply as synonymous with 'a future version of myself has experience X'.

(This is not exactly how Greaves describes her own attitude to splitting: she states that identity simply is survival, and so wishes to claim that in cases of splitting, I become each of my future selves. However, she does not make any particular effort to justify this claim other than by extension from the non-splitting case, and I think that her claim is most appropriately read simply as holding by definition: to Greaves as to Parfit, survival is what matters in all identity cases, so we might as well just use "identity" to refer to survival. Since Saunders is using the term in a very different sense, to avoid confusion I shall eschew Greaves' terminology, and simply refer to 'survival' in Parfit's manner.)

Concept failure is not, I think, something which Saunders can just reject *a priori* in the context of his thought experiment. If personal identity is an emergent concept then there is no reason why that concept should not simply break down in certain situations — especially new and alien ones, such as classical Parfittian splitting. In fact, another of Parfit's thought experiments seems to make it even more obvious that concept failure is a live option: consider a machine which merges me with Greta Garbo (Parfit 1984, pp. 229–244). Adapting the example slightly, the machine has a dial with settings from 0 to 100. Set to 0, it leaves me alone; set to 100 it obliterates me and creates Garbo *ex nihilo*; set to intermediate values it creates someone with some of my, and some of her, properties, in a ratio determined by the dial settings. I think it is hard to argue that, for settings in the vicinity of 50, there is any coherent concept of personal *identity* here, or any reasonable answer to the question 'what do you expect to happen?'.

However, once concept failure has been admitted as a conceptual possibility, it is unclear that there need be a 'correct' answer to the Saunders thought experiment. There are in fact no Parfittian splitters on Earth, so our existing concept is at best underdetermined as regards splitting, and we are actually left with the question of how (and if) it should be *extended*.

In fact, there is an extension of it available which answers Greaves' worries about the lack of uncertainty in a deterministic universe: that offered by Lewis (1976), who identifies a person (roughly) as a maximal totally ordered set of person-stages (with the ordering in question being the partial order: 'is a descendant of'). According to Lewis's proposal, if at some stage in my future I am to undergo branching into two copies, then (timelessly) there are two people, and my current (pre-branching) person-stages are shared by both of them.

On the additional assumption that the correct referent of utterances and of mental states is a person at a time (rather than a person-stage) it follows that I am genuinely ignorant of my post-branching future. For when I say 'who will I become' that statement should actually be ascribed to two versions of me (one of whom will, post-splitting, become each version of me). Since (as a consequence of any physicalist approach to the mind) any thoughts and beliefs I have at a time supervene on my person-stage at that time, and since the two versions of me share all person-stages prior to branching, it follows that it is impossible for the two versions of me to resolve their ignorance.

What are they ignorant about? Not of course any propositional knowledge, but something more indexical: something like a centred possible world (Quine 1969; Lewis 1979), but where the 'centre' is a world-line and not a point. However, presumably Greaves accepts that indexical ignorance is ignorance nonetheless, so the Lewis proposal does seem to offer an extension of our existing personal-identity concept that survives splitting.

However, just because we *can* extend our concepts in this way, it doesn't seem to be the case that we *have* to do so. In fact, I think this is a genuine choice, and one which would likely be made on sociological grounds as much as philosophical grounds.

To see this, consider another example where personal identity is in doubt: the simple (non-splitting) teletransporter, where I am disintegrated and a copy of me is assembled somewhere else from the information scanned from me in the disintegration process. Is teletransportation survival, or death followed by the creation of a doppelganger?

Well, suppose we come across an alien species who use teletransporters all the time as a form of rapid transit, and universally *believe* that teletransportation is survival. It would be hubristic (at best) to suppose that we know best here: presumably (with aliens as with other emergent objects) what justifies the validity of a given theory of personal identity is its predictive and explanatory power, and regarding an alien about to step into the teletransporter as *the same as* the one who steps out of the arrival booth is far more explanatory of the aliens' social and cultural practices.

We don't have teletransporters on Earth, so practical considerations like this aren't currently available. But suppose we did, and suppose we started using them widely; then our culture would (in that respect!) become like the aliens', and just as we did with the aliens, we should regard ourselves as surviving the teletransportation.

Would we start using them widely? I imagine that we would, but I don't know; in any case, it's a *sociological* question, and could depend wildly on extraneous factors. (Suppose, at one extreme, that a cover-up leads everyone to think that the teletransporters are really wormholes in spacetime that transport people whilst preserving their physical continuity, and people have been using them quite happily for centuries before the truth is discovered.)

It seems to me that the case of splitting is analogous. There is an extension of our theory of personal identity according to which we should expect survival and subjective uncertainty upon walking into a classical splitter, but I have no idea whether we would adopt it (I should think that it would depend very sensitively on the circumstances in which the splitters were introduced into our society).

But if the correct account of what it would be like to undergo 'classical splitting' rests on considerations like these, then — Lewis's account notwithstanding – it becomes unclear that the analogy helps us understand quantum-mechanical branching. For we should not be asking: how should we extend our concepts if branching suddenly became possible, but rather: how should we understand our existing concepts, given that branching has been happening all the time? It isn't as if we are asked what to think if we were suddenly transported from a classical to an Everettian world, or if a switch had been flipped so that Everettian branching was suddenly occurring.

For these reasons, I conclude that although I personally find Saunders' thought-experiment to be a very effective way of seeing why it *makes sense* to consider branching as subjectively indeterministic, it fails in its larger goal of showing that we are *required* to regard it thus, both because the question of how to regard classical splitting seems indeterminate and sensitively dependent on the details of its implementation in our society, and because in any case that question is not fully analogous to the question of how we should think about *quantum* branching.

3.5 Subjective Uncertainty again: arguments from interpretative charity

So: if splitting *has* been occurring all the time, how should we think about it? This brings us on to my own preferred solution to the subjective uncertainty problem (which I present more fully in Wallace (2005b)): that the problem is solved by considering how to interpret the language of inhabitants of a branching universe.

To elaborate: suppose that we consider a race of beings who inhabit a branching universe (that is, a universe like that entailed by the Everett interpretation, where one world physically splits into many), but where the beings don't realise this. Suppose further that when confronted with what are in fact branching events, they are disposed to say 'I am *uncertain what is going to occur'; (where '*uncertain') is a term in their language. More generally, suppose that they are normally disposed to assert 'A *will happen' only when it happens in every branch futurewards of the assertion, and to deny it only when it happens in no branch. However, their philosophers, asked to give an analysis of *uncertainty, are led by their ignorance of branching to the claim that one should be *uncertain of something only if there is some objective fact of the matter about which to be *uncertain, and that 'A *will happen' is true iff A happens in the single determinate future.

What are the real meanings of '*uncertain' and '*will'? One possibility (call it the Elite View) is to accept the philosophers' claims about their meaning, in which case (given that these beings' universe really branches) it appears that almost all the beings are using their language very inaccurately and are making all manner of claims which are either false or meaningless. For when the beings say (e. g.) "The Red party *will not win the election", on the Elite view they mean that in the single determinate future, the Red party do not win the election. But in their branching universe, the 'single determinate future' is one in which the world splits into many copies, in some of which the Red party win and in some of which they do not. Does this make it true that the Red party win in the future (because they do in some branches), or ill-posed that they do (because such claims presuppose falsely that there is no branching)? If the former, the beings' claim is false; if the latter, it is meaningless.

The alternative to the Elite View (call it the Charitable View) regards the beings as using the terms entirely correctly and accepts that the beings' philosophers are wrong about their language.

Which view is correct? Given plausible assumptions about the philosophy of language (notably, a certain externalism and/or holism about meaning) there can be no completely decisive answer: the beings are seriously wrong about some aspect of their world-view, but we cannot decide what they are wrong about prior to deciding their semantics. However, it is highly relevant that the Elite View makes the beings wrong almost all the time, about almost everything, whereas the Charitable View preserves the truth of most of their discourse and falsifies only fairly specialised parts of it (parts, furthermore, which were motivated by a wildly inaccurate metaphysics). According to the radical-interpretation approach to semantics espoused by Davidson (1973), Lewis(1974, 1975),Quine (1960) *et al*, there are no further facts about meaning beyond fit to usage and the best interpretation is, other things being equal, that which makes most of the community's utterances come out true. If we accept any variant of this approach, then the Charitable View easily seems to beat the Elite View.

Applying these arguments to our own language, it seems that we should conclude that our use of 'uncertain' to describe our attitude to the outcomes of quantum measurements is entirely justified. We are wrong about some of the referential underpinnings of uncertainty talk, but no wonder — the metaphysical considerations (such as the absence of widespread branching) which led us to assume those referential underpinnings were drastically wrong.[9]

(It is perhaps worth stressing that this distinction between the Elite and Charitable views is not just a linguistic dispute. Of course we can define 'uncertain' to mean whatever we like; we can define 'blancmange' to mean anything we like, too. But the argument is rather that our existing talk of future possibility and uncertainty, and the entire conceptual framework that goes with it, *already* refers to quantum branching, for all that we have not as yet realised it. As such, we are fully justified in applying our existing machinery for testing and confirming theories to the Everett interpretation, and in particular we can regard any evidence for the Quantum Algorithm as supporting the Everett Interpretation. The point will be considered more carefully in section 5.)

There is in fact a variant of the Elite View that can avoid the problems

[9] An analogy: suppose that actually the clear transparent liquid that we drink isn't H_2O at all, it's been XYZ all along, but an International Conspiracy of Chemists has hidden this from the public. Philosophers have produced semantic theories, on the basis of this faulty information, that water is necessarily H_2O. When Woodward and Bernstein uncover the Conspiracy, how will the Washington Post report it: as 'water isn't H_2O' (the Charitable View) or as 'the sea doesn't contain water' (the Elite View).

imposed by charity of interpretation.[10] Suppose (following Lewis) that in cases of an object's branching, we should regard there as having been two objects present all along, even before the branching. On this basis (as has already been discussed) an agent can maintain that he is uncertain of an outcome iff there is some fact of the matter to be uncertain about — the 'fact of the matter' is which continuant agent he is. (Or possibly: in which continuant world he is: we might more naturally apply Lewis's theory of identity to entire worlds rather than just to agents. See Wallace (2005b) for more on this matter.)

Whether we prefer to modify our metaphysics of identity or our semantics seems to me to be a question which may have no determinate answer other than utility (one is reminded of Quine's indeterminacy thesis). The conclusion is the same regardless: confronted with quantum-mechanical splitting, I should correctly assert "I am uncertain about A" whenever I know that A obtains in some but not all branches futurewards of the point of assertion.

3.6 Quantum weights and the functional definition of probability

I have argued that an agent in a branching universe should be genuinely uncertain about which outcome of branching will occur. As such, a believer in the Everett interpretation can now coherently assign credences to each possible outcome of a quantum measurement, despite his perfect objective knowledge.

Since each possible outcome is assigned a quantum-mechanical weight, we are now in a position where weight is the sort of thing that *could* fit the functional definition of probability given in section 2.1. If we simply add to the Everett interpretation the postulate that weights *in fact* fit the functional definition, we can deduce that the Everett interpretation entails the Quantum Algorithm, and as such we can regard empirical evidence for that algorithm as supporting the Everett interpretation.

If this was all that the Everett interpretation could achieve, it should still be seen as solving the measurement problem: it provides a physically complete, observer-independent theory in which is embedded the Quantum Algorithm. It may be a *postulate* that probability=weight, but the postulate is no worse off than in any other probabilistic physical theory.[11] In particular, we can perfectly well adopt the cautious functionalism espoused in section 2.3, and hope that in the future some argument will be found to justify why weight fits the functional definition.

However, things are actually rather brighter than this. There is no need for *cautious* functionalism where the Everett interpretation is concerned. As was originally argued by Deutsch (1999), and is defended in detail in Wallace (2003b) and Wallace (2003c), the principles of decision theory actually *entail* the fact

[10] Suggested in conversation by several people (most clearly by Simon Saunders) in the last few years.

[11] This position has been defended in print by Simon Saunders (1998); see also Papineau (1996).

that weight fits the functional definition. That is: in the Everett interpretation, we can prove that weight=probability.

I will not attempt to summarise these decision-theoretic proofs here, since the details are somewhat involved, but the underlying principle is essentially that of symmetry: if there is a physical symmetry between two possible outcomes there can be no reason to prefer one to another. Such arguments have frequently been advanced in non-quantum contexts but ultimately fall foul of the problem that the symmetry is broken by one outcome rather than another actually happening (leading to a requirement for probability to be introduced explicitly at the level either of the initial conditions or of the dynamics to select which one happens). They find their natural home — and succeed! — in Everettian quantum mechanics, where all outcomes occur and there is no breaking of the symmetry.

I leave it to the reader to examine these arguments in the papers cited and decide whether they are valid. If so, then the Everett interpretation has allowed us to make genuine progress on a fundamental problem in the philosophy of probability; even if not, the interpretation is no worse off than any other physical theory which makes use of objective probability.

4 Rejecting subjective uncertainty

4.1 The fission program

Notwithstanding the arguments advanced above, subjective uncertainty remains controversial. It is therefore interesting to ask to what extent we can understand the Everett interpretation without its use.

Suppose, then, that we reject subjective uncertainty. Then there are indeed no objective chances, and an agent who knows quantum mechanics (and the quantum state) is not in any way uncertain about the outcomes of measurements. Instead, such an agent knows that he has a multitude of successors: so, faced with branching, his task is to consider the interests of the (indefinitely) large number of successors which one will have after branching occurs, and to take that course of action which best serves those interests. This response (implicit in Deutsch (1999) and given explicit and elegant expression in Greaves (2004)) might be called the *fission program.*[12]It is a radical program: it entails the falsehood of a great deal of our pre-theoretic view of the world.[13] (But

[12]This is Greaves' terminology, more or less. I was tempted to call it the 'Parfittian program', but this seems a little impertinent since Parfit himself is not an advocate of it.

[13]For instance, suppose that we analyse 'untrustworthy', crudely, as 'probably isn't telling the truth'. Then, more than likely, no-one is untrustworthy (since nearly everyone is telling the truth on at least some branches), or perhaps everyone is (since nearly everyone is lying on at least some branch). A substantial fraction of the rest of our everyday concepts are similarly undermined.

Of course, the natural move is to change our analysis of 'untrustworthy': we now realise that it means 'lies on high-weight branches'. But this natural move, taken to its logical conclusion, leads back to the charity argument for subjective uncertainty, and away from the fission program.

then, given the radical nature of the Everett proposal itself, why not expect such widespread falsehood?)

The fission program can best be understood (Greaves 2004) as offering reinterpretations of the mathematical axioms of decision theory so as to apply not to an agent's ignorance of his single future but to his preferences between his multiple successors. For instance, the Dominance axiom states (roughly) that an agent should regard A as preferable to B whenever A rewards him better than B irrespective of how the future turns out (for instance, if A and B are bets where A always gives higher payoff than B, then Dominance says that A is preferable to B). The radical program reinterprets Dominance as saying that the agent should regard A as preferable to B if each of his successors is rewarded more richly under A than under B.

Each of the axioms has such a reinterpretation, and it is plausible that each reinterpretation is rationally compelling for someone in a branching universe. As such, the reinterpretation of the decision-theoretic representation theorem tells us that rational agents choose that action which maximises expected utility, where the weights in the expected-utility calculation are not credences in unknown outcomes but rather a measure of how much that agent cares about each of his determinate future descendants. Following Greaves (2004), I shall call this measure the *caring measure*.

The advocate of the fission program now proposes the following rationality principle (call it the *quantum caring principle*, or QCP): rational agents are compelled to allocate caring measure to branches in proportion to their quantum-mechanical weight, when they know the latter. That is, if E is a proposition, T is the Everett interpretation (interpreted according to the fission program) and X is the proposition that the weight of all branches on which E is true at the time in question is x (relative to the agent), then QCP requires that

$$\mathrm{Cr}(E|T\&X) = x. \tag{5}$$

If QCP is true then rational agents will act in an Everettian universe just as they would have acted in a universe where the Quantum Algorithm was true; as such, the fission program amounts to a sort of 'fictionalist' approach to the quantum algorithm, in that it entails that rational agents should behave as if there were objective probabilities even though strictly there are none.

Can we provide any sort of argument for QCP? Actually, we can provide a very good one: the decision-theoretic proofs of the Born rule mentioned in section 3.6 apply *mutatis mutandis* to the fission program under the reinterpretation of the decision-theoretic axioms, and entail that caring measure=weight. Note, though, that even if these proofs fail then QCP is not obviously worse off than the Principal Principle. That is: in both cases we appear to have a primitive rationality principle, something which we would very much rather avoid (in one case: that probability=weight; in the other, that caring measure=weight). In both cases we do not yet know how to derive that principle rather than just postulating it; in both cases we are nonetheless prepared to continue using it.

This analogy, however, is suggestive rather than conclusive. I do not see what rational argument could be given to justify our accepting the Principal Principle without argument but demanding a justification of QCP; perhaps one can be found though.

4.2 Against the fission program

Whether or not QCP is problematic, we should be slow to accept the fission program as I have so far formulated it. Partly, there are general methodological grounds to be wary of it: denying that 'uncertainty' is applicable to branching requires us to accept (given the ubiquity of branching) that most of our existing worldview is wildly wrong, in contrast with the general naturalistic viewpoint (as defended by Quine and Neurath) that progress in science and philosophy comes from successively modifying our worldview, not from rejecting it almost *in toto*.

More concretely, though, the fission program as presented above provides an answer to the wrong question. Specifically, it tells us: "supposing that we believed that the Everett interpretation was true, what would constitute rational action — that is, what rationality principles should we conform to in deciding how to live our lives?" And indeed it would be crucial to answer that question if we indeed came to believe in the Everett interpretation.

But this is not currently our situation. Rather, we want to know, 'should we believe the Everett interpretation in the first place?' That is, is the Everett interpretation explanatory of our current epistemic situation? And this, I believe, is extremely difficult to answer if we eschew all talk of uncertainty.

For recall: in section 3.1 I argued that the task of an interpretation of quantum mechanics is to embed the Quantum Algorithm that is instrumentalist QM into a satisfactory physical theory (with the possibility that it is slightly modified in the process). The fission program explicitly rejects that task when it rejects the notion of probability.

What, then, could make us come to accept the fission program? Presumably, that it offers an explanation for observed phenomena just as good as the Quantum Algorithm (while being an improvement over the Quantum Algorithm in that it is a coherent, complete scientific theory). But it is not clear why this should be so. The 'observed phenomena' are in essence a vast list of experimental outcomes whose relative frequencies correspond very closely to the probabilities defined by QM. The fission program predicts that there are branches in which this is indeed so, and ascribes a very high weight to such branches, but as yet it offers no reason why it is rational to assume that we are in one such branch. All it can provide is a prudential reason to care about successors in proportion to their weights, but that does not seem to be of epistemic import.

This point can be made more formally within the Bayesian framework for theory confirmation. Recall: in ordinary decision theory there are plausible arguments that we should update our credence in a hypothesis via conditionalising. That is, if $\mathrm{Cr}_A(B)$ is our credence in some proposition B subsequent to learning that A, then

$$\mathrm{Cr}_A(B) = \mathrm{Cr}(B|A) \equiv \mathrm{Cr}(B\&A)/\mathrm{Cr}(A). \tag{6}$$

It then follows from Bayes' Theorem that for some theory T and some evidence E

$$\mathrm{Cr}_E(T) = \mathrm{Cr}_T(E)\mathrm{Cr}(T)/\mathrm{Cr}(E). \tag{7}$$

If E is some *a priori* unlikely event assigned high objective probability by T, then (since by the Principal Principle, $\mathrm{Chance}(E) = \mathrm{Cr}_T(E)$) it follows that our credence in T will rise upon observing E.

However, according to the fission program we cannot regard the outcomes of experiments as being assigned high or low objective probability: all outcomes occur, so $\mathrm{Cr}_T(E) = 1$ irrespective of the weight of E. This seems to undermine the idea that it is our observation of *high*-weight events that provides any evidential support for quantum mechanics, since — notwithstanding QCP — the weight does not seem to appear in the Bayesian update rule.

Greaves (2004) anticipates this kind of objection, and offers a possible response. She argues that on the assumption that we live in a branching, Everettian-style universe, we can construct an analogue of the Bayesian update rule and prove its validity. But it is unclear at best how this strategy can help us where we are concerned with evidence for the Everett interpretation *itself.* For suppose that her argument succeeds. Let T be the hypothesis that the Everett interpretation is true and let X_i be the further hypothesis that the weight of branches in which evidence E occurs is x_i. Then Greaves' analogue of the update rule combined with the QCP entails that

$$\mathrm{Cr}_{ET}(X_i) = \frac{x_i \mathrm{Cr}_T(X_i)}{\sum_j x_j \mathrm{Cr}_T(X_j)}. \tag{8}$$

All that this allows us to do is to update various credences all of which are conditional on T: that is, on the truth of the Everett interpretation. It provides no way to make any statements about rational credence in the Everett interpretation itself.

The conclusion to draw from this is that we cannot assess evidence for a theory within an epistemic framework which presumes that very theory. This makes it difficult to assess evidence for Everett according to the fission program: our normal epistemic framework presumes that we are ignorant about the outcome of experiments which are to be performed, and this is simply false from the perspective of the fission program.

I am only aware of one solution to this problem which is compatible with the fission program, and (perhaps tellingly) it too makes use of the notion of ignorance. Vaidman (2002) has observed that an agent who has performed a quantum experiment but does not yet know the outcome is uncertain about that outcome in a fairly conventional way: that is, the agent knows that the outcome has some particular value but is ignorant of that value. A rational agent presumably deals with this uncertainty in the usual way, by ascribing probabilities: call these the *Vaidman probabilities* of outcomes.

A defender of the fission program can now proceed in either of two ways. The first is simply to stipulate that rational agents must set the Vaidman probability of an event equal to its quantum weight: that is, to stipulate that the quantum weights fit the functional definition of objective chance as applied to the Vaidman probabilities. The second is to argue, as above, from decision-theoretic considerations that the caring measure of an event in my future is equal to its quantum weight, and then further argue (probably by means of Dutch-Book-type considerations) that my future selves must set their Vaidman probabilities to be equal to the caring measures that I now assign to them.

In either case, the Vaidman probabilities enable us to give a conventional treatment of the epistemology of quantum mechanics. For Vaidman probability is probability nonetheless[14], and an agent who gives high probability to some measurement result E conditional on quantum mechanics will be justified in increasing his credence in the latter if he observes the former.

This strategy provides, so far as I can see, the only promising means to salvage the fission program; however, I do not find it wholly satisfactory, for two (admittedly very inconclusive) reasons. Firstly, the Vaidman probabilities are somewhat contrived entities to use as the foundation of our epistemology of quantum mechanics. It is undeniably the case that we often find ourselves in Vaidman's sort of uncertainty; however, in the bulk of experiments which we perform to test quantum mechanics the gap between our conducting the experiment and observing the result is too short to allow us time to be uncertain. (Consider, for instance, observations of a Geiger counter.)

Secondly, it is not at all clear that Vaidman probabilities can be introduced at all without admitting full-blooded subjective uncertainty. For consider: an agent about to make a measurement should expect, with certainty, that he will branch into many copies each of whom is subjectively uncertain about what result he will see when he looks. That is: he should expect, with certainty, that he will be uncertain about the result of the measurement. Is this any different from being uncertain right now about that result? I am inclined to think not, but the argument now begins to merge with the argument from interpretative charity, and I shall not pursue it further. I should stress, however, that insofar as the Vaidman probability strategy succeeds, it succeeds because it reintroduces into the Everett interpretation a notion of uncertainty, to which we can apply our existing decision and confirmation theories. Whether it is introduced via subjective uncertainty or via the Vaidman method, uncertainty of outcome result seems to be an essential component of the epistemology of the Everett interpretation.

[14] It is admittedly a slightly unusual sort of probability: not probability of being in a particular possible world, but rather probability of being in a particular location in a known possible world. This sort of *self-locating uncertainty* does seem to lead to some odd problems: see (Elga 2000) for an example.

5 Justifying the axioms of decision theory

5.1 The primitive status of the decision-theoretic axioms

At this point a sceptic might ask:

> This house of cards that you have constructed makes essential use at many points of the axioms of decision theory. What right have you to assume that those axioms hold in a branching Universe? And don't respond by reference to subjective uncertainty, please. Who is to say that the axioms apply to *this sort* of uncertainty, and not just to the more conventional sorts that Savage *et al* no doubt had in mind?

I think that there is a certain amount of force to this objection, but that as stated it misses the point. For underlying it is an epistemological story that goes like this:

> At one time, we had a metaphysical framework which included an analysed notion of uncertainty (analysed in the 'conventional'[15] way as tenseless ignorance of the state of the entire universe and/or of our location within it); at this time, though, we had no decision-theoretic axioms whatsoever. We then considered what behavioural principles would be rational for beings in our situation, and hence derived the axioms of decision theory.

But this is wildly wrong. Nothing like this actually underpins our historical propensity to conform to the decision-theoretic axioms, and in fact our reason to believe them is not based on any such argument either. Consider, for instance, why we believe an axiom like the 'Dominance' axiom mentioned in section 4.1. (Recall that it says, roughly, that an agent should regard A as preferable to B whenever A rewards him better than B irrespective of how the future turns out. Suppose we consider trying to justify the following special case:

S1 A coin is definitely going to be flipped. You have a choice of accepting a bet which pays you ten dollars if the coin lands heads, and nothing if it doesn't. You are not certain that the coin will not land heads. It is rational to take the bet.

Or consider trying to justify the following (a special case of what I call *constancy*, which is required in some versions of the decision-theoretic proofs of the quantum probability rule (Wallace 2005a)):

S2 You have to choose whether or not to have a coin flipped. You don't care whether the coin lands heads or tails (if it is flipped), and whether or not it is flipped you'll receive ten dollars. It is rational not to care about whether or not the coin is flipped.

[15] I use the term reluctantly! If the Everett interpretation is true, then branching-type uncertainty is as conventional as can be — most of our ordinary uncertainty talk refers at least in part to it.

I think the response of most people as to why they should accept these principles would be bemusement: they are, I hope, blindingly obvious. If we really pressed someone for a defence of the principles, they *might*, if they had sufficient patience, come up with something like

S1 The coin will either land heads or tails. It's at least possible that it will land heads, in which case if you accepted the bet you'll be better off. And if it lands tails, it won't make any difference. So take the bet! You won't do worse, and you might do better.

S2 If you choose to flip the coin, then it will come down either heads or tails. If it comes down heads, then you won't care whether it was flipped, because you don't care what's on the coin and you get the money anyway. If it comes down tails, likewise. So whatever happens, you won't care about whether the coin was flipped.

Note two things about these explanations. Firstly, they don't break out of the set of interconnected terms like 'will', 'might', 'uncertain', and the like. As such, if subjective uncertainty is defensible then they are just as applicable to quantum branching as to any other kind of uncertainty.

Secondly, they aren't really explanations at all, in the sense that the explanations don't involve concepts or ideas that are more basic or obvious than those used in stating the principles themselves. It's rather like trying to justify the laws of logic: if I try to justify "if (A and B), then A" by saying: "suppose 'A and B'; then in particular, A; therefore, A" I similarly haven't really explained anything, just shown some (in this case rather shallow) interconnections between equally basic concepts.

5.2 Holistic scepticism

We could conclude from the above that the decision-theoretic axioms should just be taken as primitively obvious, and left entirely unexplained; but this would be too quick. By analogy, consider the Soundness Theorem of first-order logic, which demonstrates that the rules of deduction produce only semantically valid arguments. No-one who ever doubted this fact would be convinced by the Soundness Theorem: our confidence in the laws of logic is far stronger than our confidence that we got the proof right, especially as the proof itself involves heavy use of the very logical notions which we wish to explain. Nonetheless the Soundness Theorem, showing as it does the interconnections between semantic and syntactic notions of validity, gives (some) insight into why the rules of deduction are as they are. Dummett (1978), in making this point, refers to the theorem as an *explanatory argument* for the laws of deduction; he contrasts it with *suasive arguments*, the sorts of arguments which can convince the unconvinced, and claims plausibly that no suasive argument is available here.

Further insight into the relevance of the Soundness Theorem can be gained if we imagine what we should conclude if — *per impossibile* — that theorem turned out to be false (and not false due to some isolated error, but irresolvably false).

We would not, of course, simply shrug our shoulders and stop using deduction! Rather, our entire intellectual framework would be in ruins. To countenance the possibility of such a failure would be to countenance a particularly strong form of scepticism, according to which we are not merely mistaken about many features of our world, but furthermore that world is set up so as to prevent us reasoning about it in any justified way. (I will call this 'holistic scepticism').

As another, and more naturalistic, example, suppose that we are interested in explaining the veridical nature of vision. That is, we would like to explain why are we (usually) justified in assuming that the three-dimensional world around us is as it appears to our sight. This would (conceptually, at least) be a reasonably straightforward task provided that we were studying some *other* species: a satisfactory theory of how their perceptual apparatus and their brains function will allow us to determine whether their perceptions match the outside world totally, partially, or not at all — although in the latter case the holistic framework by which we ascribe mental states to them may again start to break down and to give wildly indeterminate results.

Justifying our *own* visual capacities is more complex. We can use exactly the same scientific methodology as we might for other species, but with the proviso that we are assuming at start that our perceptions are normally accurate — else we could trust neither the readings we "observe" on our apparatus, nor the results communicated to us by co-workers. Nonetheless it seems that we can "bootstrap" our way to a satisfactory justification of our perceptual accuracy, simply because we can ourselves find the scientific explanation that third-party students of our species are able to find — and which, we have already argued, constitutes an explanation if anything does.

Furthermore, the requirement that we assume *ceteris paribus* that our perceptions are accurate does not prevent us from identifying relatively isolated flaws in that assumption, again in the same manner that third-party observers would use. Suppose for instance that everyone on Earth had vivid perceptions of winged snakes flying out of the Sun at noon. Nothing in the picture thus far developed prevents scientists from concluding that *those* perceptions are not veridical, and indeed from identifying the neural mechanisms which cause such false perceptions.

What is *not* possible in this picture is for us to identify in *ourselves* the really widespread failure of visual veridicality which we might in principle discover for another species. Such a "discovery" would so thoroughly undermine our starting assumptions as to be worthless — but in doing so it would thoroughly undermine our entire worldview. Vision, as much as deduction, is sufficiently essential to our epistemological project that it is another form of holistic scepticism to suppose that we might find that it is flawed in too widespread a way.[16]

[16]What about 'virtual reality' scenarios such as those described in films like *The Matrix* (and more soberly in some variants of the brain-in-a-vat thought-experiment)? These seem comprehensible (indeed, theoretically possible) despite the apparent widespread failure of sensory reliability which they imply. I am myself persuaded by the analysis of Chalmers (2003) (itself rather reminiscent of Putnam's (1981) treatment of the brain in the vat): no 'failure of sensory reliability' actually occurs. Rather, the various objects represented to us in

But now suppose the scenario to be modified slightly. Suppose that our argument for holistic scepticism relied on some auxiliary hypothesis H: something (let us suppose) intensely plausible, but not so conceptually indispensable as deduction or sight. Then our conclusion would be clear: H (or some other such auxiliary hypothesis) must be abandoned, however plausible it may have seen.

5.3 The role of an explanation of decision theory

Returning to decision theory, we can now see what significance might be played by an explanation of the decision-theoretic axioms. Such an explanation will have little or nothing to do with our reasons to believe decision theory, but it will give some insight into why decision theory is nonetheless correct. Furthermore, if such an explanation leads to holistic scepticism on the assumption that the world is branching, then we must abandon that assumption.

In more detail: let us suppose for a moment that the arguments of section 3 persuade us to adopt the Everett interpretation, but then we decide that our 'rational' behaviour is completely unjustified in an Everettian universe, and that the rational thing to do would be to curl up into a ball. Then the entire argument sequence — from our existing rationality, through the Everett interpretation, to the wholesale rejection of our existing rationality — would be a *reductio ad absurdum*, in effect, of whatever got us started on that sequence in the first place.

This last does not in any way suggest that the branching hypothesis is more vulnerable to such a possible criticism than the non-branching hypothesis. If we were in fact to find that it is the non-branching assumption which undermines decision theory, then branching would be forced upon us.

I have not the slightest wish to argue for such a radical and implausible conclusion; nonetheless, the non-branching case (being more familiar) provides a good starting point in any investigation of how to explain the decision-theoretic axioms. Recall that, in the non-branching picture,

1. 'Ignorance' is ignorance of the (tenseless) truth of certain facts about the (tenseless) state of the Universe. An agent, in being ignorant of some future event, simply lacks a certain item of objectively-describable knowledge.[17]

2. An agent (regarded as a person-stage) cares about future versions of himself (that is, future person-stages with the appropriate structural and causal relations to him.)

A set of defences of **S1–2** might then be:

a sufficiently all-encompassing virtual reality should genuinely be taken to exist. Our error is not in believing them to exist, but in believing them not to be computer-generated entities instantiated in some underlying hardware. (There is actually something of an analogy to the charity argument presented in section 3.5.) Further discussion, though, would take us too far afield.

[17] Actually, this simplified picture fails to take into account the need for indexical knowledge.

S1 There's something the agent doesn't know, and which he is choosing to bet on. On bet 1, his future descendant does better than on bet 2 irrespective of how that something turns out. So he isn't ignorant of which bet will be better for his descendant, even though he is ignorant about the outcome of the bet.

S2 The agent can make one of two bets, and whichever bet he makes his future descendant receives a a certain fixed reward. According to the bet he chooses, that descendant's environment has certain properties about which he is ignorant, but to which he is indifferent. Therefore, the agent himself should be indifferent as to those properties; therefore, he should not care which bet he takes.

I hope that it is clear that no-one comes to believe $S1$ or $S2$ on the basis of arguments such as these! They are explanatory only in the sense that they help us come to a better understanding of why the principles are true, not in the sense that they would convince someone who did not already accept them that they should reconsider.

Can we construct analogous arguments in the case of branching? It is here, I think, that "Parfittian" arguments about caring for our successors, such as those involved in the 'fission program', come into place. Recall, for instance, the fission-program reinterpretation of Dominance: an agent is justified in preferring action A over action B iff all his successors do better under A than under B. If this is regarded not as a *reinterpretation* of the Dominance axiom but as an *explanation* of it, it is (I claim) every bit as reasonable an explanation as the analogous non-branching explanation of Dominance.

Arguments such as these aim to establish, in effect, that even if we knew that future-directed uncertainty was really branching, we should rationally continue to behave exactly as we did before we knew this. Our existing attitude to decision-making in the face of branching just *is* to treat it as uncertainty, because that's what words like 'uncertainty' actually refer to if the Everett interpretation is true. But the arguments of the fission program show us why the decision-theoretic axioms are justified on the assumption that it is true.

This is not to deny that *some* isolated aspects of our view of rationality might need revision if we accept the Everett interpretation (rather as the 'winged snakes' of the vision example above would imply a localised revision of our assumption that vision is reliable). One possible example, long discussed informally among physicists and recently analysed by Lewis (2004), is *quantum Russian roulette*: bet a large sum on an unlikely quantum event, and arrange to be instantly obliterated if you lose. It has been argued that if the Everett interpretation is correct then it is rational to expect survival with certainty in these experiments, and thus to agree to partake in them. (I'm sceptical myself, but will not defend that scepticism here.) As always, we sail in Neurath's boat, and no part of our conceptual scheme is completely insulated from criticism and revision. But it *is* to deny that such revisions could be so widespread as to undermine our original reasons for coming to accept the Everett interpretation.

We seem to find that the epistemology of the Everett interpretation has a two-part structure. First we must ask: is the interpretation explanatory of our current epistemic situation, on the assumption that our existing approaches to decision-making and uncertainty are basically rational? If the answer is yes, we must also ask (on pain of holistic scepticism) whether that assumption remains valid if the Everett interpretation is indeed true. I claim that we are now in a position to give affirmative answers to both questions.

6 Conclusions

Probability enters our scientific theories through the Principal Principle and only through the Principal Principle. As such, our term 'objective probability', if it picks out anything, picks out that physical property which satisfies the functional definition given implicitly by the Principle. Since regarding that functional definition as true by postulate (i. e. , primitivism) is deeply unattractive, the best attitude to probability is a cautious functionalism whereby we assume that something can be proved to satisfy the definition even though we do not currently have any good candidates.

A solution of the measurement problem is an embedding of the 'quantum algorithm' — that is, the algorithm by which we calculate the objective probabilities of outcomes to experiments — within a complete, physically satisfactory theory. The Everett interpretation is such a solution[18] provided that the problem of 'subjective uncertainty' can be solved: that is, provided that a way can be found to justify an agent who believes the Everett interpretation nonetheless being uncertain about the outcomes of a measurement. However, solutions to this problem are available via interpretative charity and Lewisian treatments of identity (and possibly also via post-measurement ignorance, i. e. Vaidman probability).

If a solution to the subjective uncertainty problem can indeed be found, then the "weights" of branches are candidates for probability in the sense that they are the right sorts of properties to fit into the functional definition of probability offered by the Principal Principle. If this were all that could be said, the Everett interpretation would be no worse off than any other physical theory involving probability, since no such theory has any argument for why its 'probability' fits the functional definition.

However, it is not all that can be said. Rather, the axioms of decision theory combined with the mathematical structure of quantum mechanics suffice to derive the Principal Principle with weights playing the roles of probability; that is, according to the functional definition of probability, we can prove that weights are probabilities.

An alternative approach to the Everett interpretation (which I have called the 'fission program') makes no use of subjective uncertainty, but instead reinterprets the axioms of decision theory as axioms about how an agent should care about his successors in the case of branching. This program must be rejected in

[18]On the assumption that decoherence solves the preferred basis problem.

its full-strength form, as it does not provide an epistemically acceptable account of how we can come to accept the Everett interpretation. It does, however, have an important role to play in the epistemology of the Everett interpretation: after we have tentatively come to believe the interpretation, the fission-program reinterpretations of the axioms serve as 'explanatory arguments' for the validity of those axioms, forestalling worries that the Everett interpretation must be rejected because it undermines our overall conceptual scheme too drastically.

A summary of the epistemic route to the Everett interpretation — and, perhaps, of the route to any substantial revision of our conceptual scheme — might be: first see whether your existing machinery for theory appraisal recommends that you adopt the new theory. If it does, see whether that 'existing machinery' is still essentially valid after adopting the new theory's viewpoint. If the first step fails, the theory is straightforwardly to be rejected; if the second fails, the reasons for rejecting the theory are more subtle but no less pressing. Happily, it appears that neither fails in the case of the Everett interpretation: it solves the measurement problem in a fully satisfactory way.

Acknowledgements

I am grateful to all the participants in recent discussions of these matters in Oxford and via email, but in particular to Hilary Greaves, Wayne Myrvold and Simon Saunders. I am also grateful to Jeremy Butterfield for his many constructive comments on an early version of this paper.

References

Chalmers, D. (2003). The Matrix as metaphysics. Available online at http://consc.net/papers/matrix.html.

Davidson, D. (1973). Radical interpretation. *Dialectica 27*, 313–328. Reprinted in Donald Davidson, *Enquiries into Truth and Interpretation* (Oxford University Press, Oxford, 1984).

Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin.

Deutsch, D. (1999). Quantum Theory of Probability and Decisions. *Proceedings of the Royal Society of London A455*, 3129–3137. Available online at http://arxiv.org/abs/quant-ph/9906015.

Dummett, M. (1978). The justification of deduction. In *Truth and Other Enigmas*. London: Duckworth.

Elga, A. (2000). Self-locating belief and the sleeping beauty problem. *Analysis 60*(2), 143–147.

Fuchs, C. (2002). Quantum mechanics as quantum information (and only a little more). Available online at at http://arXiv.org/abs/quant-ph/0205039.

Fuchs, C. and A. Peres (2000). Quantum theory needs no "interpretation". *Physics Today 53*(3), 70–71.

Greaves, H. (2004). Understanding Deutsch's probability in a deterministic multiverse. *Studies in the History and Philosophy of Modern Physics 35*, 423–456. Available online at http://arXiv.org/abs/quant-ph/0312136 or from http://philsci-archive.pitt.edu.

Jeffrey, R. C. (1983). *The Logic of Decision (2nd edition)*. Chicago: University of Chicago Press.

Joyce, J. N. (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Kaplan, M. (1996). *Decision Theory as Philosophy*. Cambridge: Cambridge University Press.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy 67*, 427–446. Reprinted in David Lewis, *Philosophical Papers*, Volume I (Oxford University Press, Oxford, 1983).

Lewis, D. (1974). Radical interpretation. *Synthese 23*, 331–344. Reprinted in David Lewis, *Philosophical Papers*, Volume I (Oxford University Press, Oxford, 1983).

Lewis, D. (1975). Languages and language. In K. Gunderson (Ed.), *Minnesota Studies in the Philosophy of Science*, Volume VII. Minnesota: Minnesota University Press. Reprinted in David Lewis, *Philosophical Papers*, Volume I (Oxford University Press, Oxford, 1983).

Lewis, D. (1976). Survival and identity. In *The Identities of Persons*. Berkeley: University of California Press. Reprinted in David Lewis, *Philosophical Papers*, Volume I (Oxford University Press, Oxford, 1983).

Lewis, D. (1979). Attitudes *De Dicto* and *De Se*. *Philosophical Review 88*, 513–543. Reprinted in David Lewis, *Philosophical Papers*, Volume II (Oxford University Press, Oxford, 1986).

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, Volume II. Berkeley: University of California Press. Reprinted in David Lewis, *Philosophical Papers*, Volume II (Oxford University Press, Oxford, 1986).

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*, 343–377. Reprinted in David Lewis, *Papers in Metaphysics and Epistemology* (Cambridge: CUP, 1999), pp. 8–55.

Lewis, D. (1986). *Philosophical Papers, Vol. II.* Oxford: Oxford University Press.

Lewis, D. (1994). Chance and credence: Humean supervenience debugged. *Mind 103*, 473–90. Reprinted in David Lewis, *Papers in Metaphysics and Epistemology* (Cambridge: CUP, 1999), pp. 224–247.

Lewis, D. (2004). How many lives has Schrödinger's cat? In F. Jackson (Ed.), *Lewisian Themes: the philosophy of David K. Lewis*, pp. 4–23. Oxford: Oxford University Press.

Mellor, D. (1991). Properties and predicates. In *Matters of Metaphysics*, pp. 170–182. Cambridge: Cambridge University Press. Reprinted in *Properties*, D. H. Mellor and Alex Oliver (eds.) (Oxford: OUP, 1997), pp. 255–267.

Mellor, D. H. (1971). *The Matter of Chance.* Cambridge: Cambridge University Press.

Papineau, D. (1996). Many Minds are No Worse than One. *British Journal for the Philosophy of Science 47*, 233–241.

Parfit, D. (1984). *Reasons and Persons.* Oxford: Oxford University Press.

Putnam, H. (1981). *Reason, Truth and History.* Cambridge: Cambridge University Press.

Quine, W. V. (1969). *Propositional Objects*, pp. 139–160. New York: Columbia University Press.

Quine, W. v. O. (1960). *Word and Object.* Cambridge, Mass.: MIT Press.

Saunders, S. (1998). Time, Quantum Mechanics, and Probability. *Synthese 114*, 373–404.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover.

Shoemaker, S. (1980). Causality and properties. In P. van Inwagen (Ed.), *Time and Cause: essays presented to Richard Taylor*, pp. 109–135. Dordrecht: Reidel. Reprinted in *Properties*, D. H. Mellor and Alex Oliver (eds.) (Oxford: OUP, 1997), pp. 228–254.

Vaidman, L. (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. Available online at http://arxiv.org/abs/quant-ph/9609006.

Vaidman, L. (2002). The Many-Worlds Interpretation of Quantum Mechanics. The Stanford Encyclopedia of Philosophy(Summer 2002 Edition), Edward N. Zalta (ed.), URL=http://plato.stanford.edu/archives/sum2002/entries/qm-manyworlds.

Wallace, D. (2003a). Everett and Structure. *Studies in the History and Philosophy of Modern Physics 34*, 87–105. Available online at http://arXiv.org/abs/quant-ph/0107144 or from http://philsci-archive.pitt.edu.

Wallace, D. (2003b). Everettian rationality: defending Deutsch's approach to probability in the Everett interpretation. *Studies in the History and Philosophy of Modern Physics 34*, 415–439. Available online at http://arXiv.org/abs/quant-ph/0303050 or from http://philsci-archive.pitt.edu.

Wallace, D. (2003c). Quantum probability from subjective likelihood: Improving on Deutsch's proof of the probability rule. Available online at http://arXiv.org/abs/quant-ph/0312157 or from http://philsci-archive.pitt.edu.

Wallace, D. (2005a). Probability in three kinds of branching universe. In preparation.

Wallace, D. (2005b). Tensed talk in a branching universe. Forthcoming.