A formal proof of the Born rule from decision-theoretic assumptions

David Wallace

June 15, 2009

Abstract

I develop the decision-theoretic approach to quantum probability, originally proposed by David Deutsch, into a mathematically rigorous proof of the Born rule in (Everett-interpreted) quantum mechanics. I sketch the argument informally, then prove it formally, and lastly consider a number of proposed "counter-examples" to show exactly which premises of the argument they violate.

(This is a preliminary version of a chapter to appear — under the title "How to prove the Born Rule" — in Saunders, Barrett, Kent and Wallace, *Many worlds? Everett, quantum theory and reality*, forthcoming from Oxford University Press.)

1 Introduction

> Thus we see that quantum theory permits what philosophy would hitherto have regarded as a formal impossibility, akin to "deriving an ought from an is", namely deriving a probability statement from a factual statement. This could be called deriving a "tends to" from a "does". (Deutsch 1999)

The "Everett interpretation of quantum mechanics" is just unitary quantum mechanics, taken literally as a description of the world; it is a "many-worlds" theory because it instantiates multiple, emergent, branching quasiclassical realities. That much is commonplace amongst contemporary Everettians; it was argued for *in extenso* in my other chapter in this volume, and for the purposes of *this* chapter I will take it as read.

It is widely held, however, that a problem remains: namely, how does *probability* fit into this story? It is not in dispute what physical magnitude is *supposed* to be (or to stand in for) probability: the probability of a branch is supposed to be its weight (i. e., its mod-squared amplitude). More formally, the probability of a history α represented by a history operator $\widehat{C}_\alpha$ (in the consistent-histories formalism) is supposed to be

$$\Pr(\alpha) = \langle\psi|\, \widehat{C}_\alpha \,|\psi\rangle \tag{1}$$

where $|\psi\rangle$ is the universal (Heisenberg-picture) state. In more parochial language, if an observer's branch has weight w_0, if it is going to split into multiple branches, and if those branches in which X happens have weight w_X, then the probability for that observer of X happening is supposed to be w_X/w_0.

What is in dispute is why, and how, this physical magnitude can be probability. One might ask: how can it even *make sense* for anything to "be probability" in a theory where all possible outcomes occur; less confrontationally, one might ask what kind of *argument* can be given to justify the claim that mod-squared amplitude is probability. (I have previously referred to these as the *Incoherence Problem* and the *Quantitative Problem* respectively, though I shall not make much of the distinction in this article.)

One quite legitimate response to this question, I think, is bemusement. After all, formally speaking the measure defined by mod-squared amplitude on any given space of consistent histories satisfies the algorithms for a probability. Indeed, mathematically the setup is identical to any stochastic physical theory, which ultimately is specified by a measure on a space of kinematically possible histories (albeit that measure is usually given indirectly, via a stochastic differential equation). In Everett-interpreted quantum mechanics, the correct space of consistent histories is the space of quasi-classical histories; as my earlier chapter argued, this space may be imprecisely defined but this is no reason not to take it seriously as emergent structure. So (goes the response) all the formal requirements to take mod-squared amplitude as probability are in place; to ask for more is no more justified than to ask why the physical quantity represented by the metric in Newtonian space "is" length, or why other mathematical features of classical physics represent mass or charge. (Saunders (1998) develops this response in more depth.)

I have a good deal of sympathy for this response, but in this chapter I wish to discuss a more positive answer to the question, which might be called the *decision-theoretic strategy.*

Decision-theoretic strategy: Probability gets its meaning in quantum mechanics through the rational preferences of agents. In particular, a rational agent who knows that the Born-rule weight of an outcome is p is rationally compelled to act as if that outcome had probability p.

The decision-theoretic strategy was advocated by Deutsch (1999). In the same paper he presented an informal proof, from principles of decision theory that (he argued) did not themselves invoke probabilistic notions, that the only rational strategy for an agent in an Everettian universe is to follow the Born rule. I developed Deutsch's proof further, and presented alternative versions of it, in Wallace (2003) and Wallace (2007).

The argument has met with its share of criticism. Some of the criticism (e. g. Barnum *et al* (2000), Lewis (2005)) has been directed at the proof itself; some (e. g. Albert and Price's contributions to this volume) has sought to undermine the possibility of a proof by proposing other, (allegedly) equally rationally justifiable alternatives to the Born rule. (These, I should make clear, are not proposed as positive *suggestions* as to how we should act if the Everett interpretation is correct; they are intended rather as *reductios*.)

My purpose in this chapter is to give a self-contained defence of the decision-theoretic strategy, culminating in a formal proof of the rational necessity of the Born rule from axioms of decision theory which I will defend. Formalisation is not often an aid to understanding, but when a result is controversial it can be helpful to see *exactly* what is and is not required. Along the way I will showcase some of the various proposed alternative strategies for Everettian rationality, and show exactly how they conflict with the assumptions in the argument; in doing so, I hope, the argument for making those assumptions will become clearer.

My focus in the chapter is deliberately narrow. A proof that rational agents in an Everett universe must act in accordance with the Born-rule probabilities falls short of a full solution of the probability problem: we might also ask how this decision-theoretic notion of probability connects with our use of probability in assessing the evidence for quantum mechanics, or with our ordinary, pre-theoretic notions of probability as a guide to action in cases of uncertainty. I shall address neither question here, though (for discussions of the former question, see Greaves and Myrvold's contribution to this volume, Wallace (2006), and part II of Wallace (2010); for the latter, see Wallace (2005), part III of Wallace (2010), Saunders (1998), and Saunders' contribution to this volume.)

I shall begin in section 2 with a brief discussion of decision theory *in general*, and go on in section 3 to see how decision theory works in the Everett interpretation. The main part of the paper (sections 4–8) states and proves, first informally and then in full mathematical rigor, that the Born rule is the unique rational strategy available in the Everett interpretation. I illustrate this result (section 9) by considering a number of other proposed strategies, proposed at various times and places as counterexamples to the necessity of the Born rule, and show why those strategies are not in fact valid alternatives. I conclude with a few more general observations about why the Born rule can indeed be proven, and why the Everett interpretation is essential in such a proof.

2 Preamble: the decision-theoretic approach

Suppose an coin is to be tossed in five minutes' time; and suppose that an agent bets five dollars (at even odds) that it will land heads. There are two interestingly different possible results: (i) the measurement gives result "up" and the agent gets \$5; (ii) the measurement gives result "down" and the agent loses \$5. If the result is heads, the agent will be pleased about the bet; if it is 'tails he will be less delighted, though (if he is of an appropriate character) he may well still regard the bet as having been the right choice given his information before the coin toss. (There are of course vastly more than two microscopically distinct possible results; the division into two sets is based on the pragmatic interests of the agent.)

In deciding whether to accept this bet as opposed to any number of other bets, the agent has to weigh the cost to him if the result is "down" against the benefit if it is "up". Decision theory gives a precise answer to the question of how he should carry out this weighting: he should assign some utility (some real number) $\mathcal{V}(+\$5)$ to receiving five dollars, and some other utility $\mathcal{V}(-\$5)$ to losing five dollars, and some third utility $\mathcal{V}(\$0)$ to neither getting nor losing it; and he should assign a probability $\Pr(H)$ to heads; and he should take the bet only if

$$\Pr(H) \times \mathcal{V}(+\$5) + (1 - \Pr(H)) \times \mathcal{V}(-\$5) > \mathcal{V}(\$0). \tag{2}$$

More generally, decision theory mandates that an agent should assign a utility to each payoff, and a probability to each outcome, and that faced with *any* decision, the agent should choose that option which maximises expected utility with respect to those assignments.

(In elementary discussions, it is common just to assume $\mathcal{V}(+\$N) = N$, but this is too simplistic: it is not irrational to refuse to trade your house for a one-in-a-thousand chance of winning Microsoft. In fact, in a decision-theoretic framework, what it *means* to say that one reward is twice as valuable as another is that a 50% chance of getting the first is as valuable as getting the second with certainty. See Savage (1972, pp.91–104) for more on this point.)

Why should an agent behave this way? Prima facie, it isn't obvious at all that he should try to maximise expected utility rather than, say, maximising utility with respect to the square of the probability function; or maximising the logarithm of utility; or just maximising the utility of the least good possible outcome.

Decision theory has a standard answer to this question: if an agent has a definite preference (which might be indifference) between any two bets, and if that preference order obeys certain constraints which are purported to be necessary conditions of rationality, and if the set of available bets has a sufficiently rich structure, then it is possible to prove a *representation theorem*: a theorem that for any such preference order there is a unique probability function, and an essentially unique[1] utility function, such that one bet is preferred to another iff it has a higher expected utility. It follows that any agent whose preferences cannot so be represented must be acting irrationally: that is, must somewhere be violating a principle which (again, purportedly) is a necessary constraint on rational action.

It will be instructive to present two such principles (both drawn from Savage (1972)). The first is *transitivity*: if an agent prefers a to b and b to c, he should prefer a to c. The second might be called *dominance*: if an agent will do better through a than b *whatever happens*, he should chose a over b. (For instance, a bet that pays ten dollars if a coin-toss lands heads and nothing if it lands tails is to be preferred over a bet on the same toss that pays five dollars on heads and minus five on tails).

There is an important weakness in this decision-theoretic argument which

[1] By "essentially unique" I mean unique up to positive affine transforms $x \longrightarrow ax + b$, with a positive. Fairly obviously, such transformations serve only to scale the expected utility of all bets in the same way.

needs to be stressed. It is a proof that rational agents must bet according to *some* probability function, but it is silent on the connection between that function and the "real" probabilities. No decision-theoretic principle is contradicted by an agent who assigns probability 99/100 to an apparently-fair coin landing "heads", for instance. A minority of advocates of the decision-theoretic approach simply deny that there is any such thing as objective or "real" probability; the majority just take it as a bare postulate that an agent should conform his subjective probabilities to the objective probabilities when he knows the latter.[2]

This weakness actually rather undermines the use of probability as a criticism of the Everett interpretation, even without the arguments of this paper: if classical probability can give no justification of its probability rule, why ask the Everettian for such a justification? But in fact, we will see that in the Everett interpretation, not only can we use rationality considerations to make sense of probabilities as well as in conventional decision theory, we can prove and not merely postulate the link between those probabilities and the quantum-mechanical weights.

3 Everettian Rationality

So: consider the Everettian version of the coin-toss. Instead of a coin, we have a particle in a superposition of spin-up and spin-down (in some fixed direction); instead of a coin-toss, we have a spin measurement. And instead of there being two interestingly different *possibilities*, there are two interestingly different sets of branches: the spin up branches, where (if the agent took the bet) he gets five dollars, and the spin down ones, where he loses five dollars. In deciding whether to accept this bet as opposed to any number of other bets, the agent has to weigh the benefit to himself in the branches where the result is up against the cost in the branches where the result is down. So the notion of a bet is at least meaningful in the Everettian context.

Skeptic: The benefit isn't to the agent: it's to copies of that agent in the future.

Author: Sure. But that's true in the non-Everettian just as in the Everettian case. In either case, the reason the agent makes one choice rather than another is because of his concern about his future interests — that is, about the interests of his future self or selves.

And an agent's future self *is* his future self just by virtue of the causal, structural, dynamical relations between it and the agent's past self. There is (I assume!) no indivisible, immaterial soul which passes through my life and magically makes me a single being: what makes the stages of me at different times all *me* is that they are appropriately related. And it

[2] This is a simplification: a more general statement is that an agent's subjective probability in X conditional on the objective probability of X being p should in turn be p (this is known as the Principal Principle, following Lewis (1980)). My main point stands, however: the Principal Principle is postulated, not derived.

seems, at least, that an Everettian agent's future selves stand in all the same relations to him as a non-Everettian agent's future selves stand to *him.*

Skeptic: There's a pretty obvious disanalogy, though. In the Everettian case, there's more than one future self!

Author: Fair enough. (Though there are some subtleties here: see Saunders and Wallace (2008), and also Saunders' chapter in this volume, for a construal of personal identity in which this is not the case.) But it's hard to see why, in the Everett case, I should regard my future self as any the less *me* — why I should not treat his goals and desires, his hopes and dreams, as my own — just because I actually have multiple such selves. And so it's hard to see why those future selves should be less relevant to my considerations now — to my decision-theoretic preferences — than would be the case in the absence of Everettian branching.

So an Everettian agent can be in a decision problem — can be faced with a choice of bets — just as can a non-Everettian. And certainly *one* strategy available to him is what might be called the "Born-rule strategy": choose that bet which maximises expected utility with respect to the Born-rule weights (that is, the mod-squared amplitudes). An Everettian agent who adopted the Born-rule strategy would make exactly the same choices between bets as would a non-Everettian who adopted the Principal Principle with respect to the Born-rule weights. The two would be indistinguishable in terms of their behavioural dispositions.

Is it the *only* strategy? Advocates of Deutsch's decision-theoretic strategy say that it is. More precisely, they argue that given certain principles of rationality, and given knowledge of quantum mechanics, it can be proved that any strategy other than the Born-rule strategy violates some rational constraint on action. By analogy with the Representation Theorems of classical decision theory, we might call such a result a (purported) *Quantum Representation Theorem.*

Such a theorem can in fact be proved formally, and in sections 7-8 I will give a formal proof of such a theorem. But since a fully formalised proof of a result does not make for accessible reading, firstly I will give an *informal* version. In sections 4-5 I will state informally, and motivate, the axioms I wish to use; in section 7 I will argue informally why these axioms jointly entail a Quantum Representation Theorem.

4 The quantum decision problem

The situation I wish to consider is the following. A quantum state is to be prepared in some superposition; the system is measured in some basis; a bet is made by the agent on the outcome of that measurement. Our agent knows (we assume) that the Everett interpretation is correct; he is also assumed to know the universal quantum state, or at least the state of his branch. (The latter is an unrealistic but convenient assumption; in practice, however, it suffices for the agent to know the mod-squared amplitudes for each outcome of a measurement.) His preferences can be represented by an ordering relation on these bets.

Since (in Everettian quantum mechanics, at any rate) preparations, measurements, and payments made to agents are all physical processes, there is a certain simplification available: any preparation-followed-by-measurement-followed-by-payments can be represented by a single unitary transformation. So our agent's rational preference is actually representable by an ordering on unitary transformations.

We should acknowledge that not all unitary transformations represent something physically possible.[3] In particular, transformations which lead to *recoherence* — that is, to Everett branches merging — are certainly not performable by any agent localised to a specific branch. But nonetheless we will consider a fairly wide set of transformations to be available — exactly how wide is something that the axioms will spell out. (It might be worth recalling at this stage that decision theory is concerned with the preferences an agent would have when confronted with a particular decision — his dispositional preferences, in philosophers' language — and not just with what actually happens. It is most unlikely that I will be offered a choice between the presidency of the World Bank and the deputy leadership of the Al-Qaeda terror group, but I have a definite preference between the two. As such, the assumption of a reasonably wide set of transformations seems reasonable enough.)

We should also acknowledge in our decision-theoretic setup that decoherence imposes a certain structure on the Hilbert space. We can represent this by a resolution of the identity on the Hilbert space: that is, by a decomposition of the space into subspaces, with each subspace π corresponding to a possible macrostate. The choice of macrostates is largely fixed by decoherence, although the precise fineness of the grain of the decomposition is underspecified. (In the model, of course, it will be precisely specified, but this just illustrates that the model is artificially precise.) We call a macrostate *available* to an agent if there is an available act which, when performed, leaves some of his future selves in that macrostate.

Part of the point of the decomposition into macrostates is that an agent can be assumed not to care exactly what the microstate is within a given macrostate (if he does care, we have defined the macrostates too coarsely). But in fact, usually an agent will also be indifferent between a great many macrostates: for instance, if offered a million dollars, I am indifferent as to the colour of the cheque.[4] It will be useful to consider a coarse-graining of the macrostate

[3]Doesn't *only one* unitary transformation represent something physically possible? Doesn't the Hamiltonian of the universe uniquely determine which transformation is performed? If this is a problem, it is not specific to Everett: it is the ancient debate of free will vs. determinism. Rather than get into this morass (though I recommend Dennett (1984) for reassurance that the two are compatible), let me just note that we can talk about *rational strategies* even if an individual agent is not free to choose whether or not his strategy is rational.

[4]The reader who doubts this claim is encouraged to test it empirically.

subspaces into *reward subspaces*, such that an agent's only preference is to which reward subspace he is in. Formally speaking, "reward subspace" is a derived concept within the decision theory.

In fact, for mathematical reasons it will be convenient to work both with the set of macrostates and with the Boolean algebra $\mathcal{E}$ of arbitrary disjunctions of macrostates[5], which we call the *event space*. The formal development of the theory will not actually require the assumption that the event space can be constructed from a set of macrostates (though it does not rule out that assumption). Indeed, since the fineness of grain of branches is indeed underspecified, the branch structure might be best idealised in some particular situation by a model in which the algebra is not constructed this way. (For instance, if the Hilbert space is $L^2(R^N) \otimes \mathcal{H}_E$, where $\mathcal{H}_E$ represents some subsystem of environmental degrees of freedom, then we might wish to take the elements of $\mathcal{E}$ to be the subspaces

$$\Sigma_E = \{f \otimes v : E \text{ is an open subspace of } R^N \text{ and } f \text{ has support in E}\} \tag{3}$$

which cannot be generated from macrostates (unless we are willing to relax rigor and consider eigenstates of position).

For simplicity, we will refer to the set of unitary transformations over which an agent's preference order is defined as *acts*. A different set may be relevant for different physical states of the universe, so we will have cause to speak of the *acts available at* a macrostate π. (In view of the previous paragraph's comment, we might do better to talk of the acts which are *contemplatable* at π; I avoid this terminology mostly because it's cumbersome.)

In fact, it will be simpler to talk of which acts are available at a given event (not just a macrostate) — informally an act available at an event $E = \pi_1 \vee \pi_2 \vee \cdots \vee \pi_N$ is the conditional act 'if the macrostate is actually π_i, perform U_i. This makes it much more straightforward to talk about the composition of acts: if U is available at an event E, and V is available at the smallest event containing the range of V, for instance, then VU ought to be the act of performing U and V sequentially and so also should be available at E. In the formal development we will state explicit rules to ensure that these and similar compositions are available; for now we take it as tacit that they are.

We now need to represent the agent's preferences between acts. Since those preferences may well depend on the state, we write it as follows: if the agent prefers (at ψ) act U to act U', we write

$$\widehat{U} \succ^{\psi} \widehat{U}'. \tag{4}$$

To be meaningful, of course, this requires that U and U' are both available at ψ's macrostate. So $\succ^{\psi}$ is to be a two-place relation on the set of acts available at that macrostate. In the event formalism we use later, we will require $\succ^{\psi}$ to be a two place relation on the acts available at each event which contains ψ.

[5]Recall that the disjunction $E \vee F$ of two subspaces of a Hilbert space is the closure of the span of their union.

So much for the setup; now for the axioms. They come in two categories: axioms of *richness*, which concern which acts are available to the agent (how rich the structure of the set of acts is) and which are not connected to a particular agent's preference order; and axioms of *rationality*, which constrain that preference order.

The richness axioms, then, are

Reward availability: All rewards are available to the agent at any macrostate: that is, the set of available acts always includes ones which give all of the agent's future selves the reward.

Branching availability: Given any set of positive real numbers $p_1, \ldots p_n$ summing to unity, an agent can always choose some act which has n different macrostates as possible outcomes, and gives weight p_i to the ith outcome.

Erasure: Given a pair of states $\psi \in E$ and $\varphi \in F$ in the same reward, there is an act $\widehat{U}$ available at E and an act $\widehat{V}$ available at F such that $\widehat{U}\psi = \widehat{V}\varphi$.

Problem continuity: For each event E, the set of acts available at E is an open subset of the set of unitary transformations from E to $\mathcal{H}$.

These should mostly be uncontroversial. Branching availability and reward availability are consequences of the relatively stylised decision problem we are considering, where measurements are being made and payments are being provided; they reflect the facts (respectively) that quantum systems can be prepared in arbitrary states and that envelopes of cash can always be given to people.

Erasure is slightly more complicated. It effectively guarantees that an agent can just forget any facts about his situation that don't concern things he cares about (that, is, by definition: that don't concern where in the reward space he is). In thinking about it, it helps to assume that any reward space has an "erasure subspace" available (whose states correspond to the agent throwing the preparation system away after receiving the payoff but without recording the actual result of the measurement, say). An "erasure act" is then an act which takes the quantum state of the agent's branch into the erasure subspace; the agent is (by construction) indifferent to performing any erasure act, and since he lacks the fine control to know which act he is performing, all erasures should be counted as available if any are. It follows that, since for any two such agents all erasures are available, in particular there will be two erasures available satisfying the axiom.

I postpone a discussion of problem continuity until the axioms of rationality have been introduced.

5 The dictates of rationality

Moving on to the rationality axioms, they come in two groups. The first two axioms are very general principles of rationality, as relevant in the classical as in the quantum context.

Ordering: The relation $\succeq^{\psi}$ is a total ordering for each ψ on the set of acts available at ψ, for each ψ (that is: it is transitive, irreflexive and asymmetric, and if we define $U \sim^{\psi} V$ as holding whenever $U \succ^{\psi} V$ and $V \succ^{\psi} U$ fail to hold, then $\sim^{\psi}$ is an equivalence relation).

Diachronic consistency: If U is available at ψ, and (for each i) if in the ith branch after U is performed there are acts V_i, V_i' available, and (again for each i) if the agent's future self in the ith branch will prefer V_i to V_i', then the agent prefers performing U followed by the V_is to performing U followed by the V_i's.

Ordering is utterly familiar (indeed, built in to our use of the $\succ^{\psi}$ symbol) and hopefully uncontroversial. But it is worth stressing that the *reason* it is uncontroversial is not (just!) that it would be unintuitive for an agent's preferences to violate ordering, but because it isn't even possible, in general for an agent to formulate and act upon a coherent set of preferences violating ordering.

Of course, in stylised and artificial special cases, it might be. If an agent knows that he will be offered three acts chosen from a set of ten, he can arbitrarily pick one element from each three-element subset, and elect to choose that one. But of course, real decision problems aren't that cleanly specified: the precise number of acts available is vague or just indeterminate and the cognitive cost of trying to pin down the size of that set is prohibitive (even when the very act of trying to pin it down does not change the problem out of recognition). Excluding stylised and occasional exceptions, then, ordering is *constitutive* of rationality, not just intuitively necessary for it.

I have stressed this because, in fact, very much the same defence can be offered of the less-familiar diachronic consistency principle, which in effect rules out the possibility of a conflict of interest between an agent and his future selves. In philosophy examples one often speaks of a (classical) agent as if he were a continuum of independent entities, one for each time, each having his own preference ordering. But of course actual decision-making takes place over time. An agent's actions take time to carry out; his desires and goals take time to be realised. If his preferences do not remain consistent over this timescale, deliberative action is not possible at all.

Of course, there are plenty of *localised* violations of diachronic consistency even outside the Everettian context. If I tell my friend not to let me order another glass of wine after my second, I acknowledge that my desires at that point will conflict with my desires now. But notice that such situations

(a) are generally not taken to be rational;

(b) are indeed analysed as situations of conflict, where my present self acts to prevent my future self having access to his preferred choice;

(c) are localised, taking place against a general assumption of diachronic consistency in myself and others (as when I assume that my friend's future self will indeed act on her agreement not to let me order the wine, or that the morning after the night before, I'll be glad that she did).[6]

Similarly: in a branching universe, to accept a conflict of interest between my pre-branch and post-branch selves is to cease to see them as the same person. If branching were an isolated occurrence, this might be possible: it is arguably callous to make a copy of myself and send him off to do a dangerous or disagreeable task — and, crucially for the point, to take actions designed to prevent him shirking that task but it is not *irrational*.[7] But *Everettian* branching is ubiquitous: agents branch all the time (trillions of times per second at least, though really any count is arbitrary). In the presence of *widespread, generic* violation of diachronic consistency, agency in the Everett universe is not possible at all.

Skeptic: Stop there. You're trying to argue that rationality (agency, if you like) even makes sense in an Everett universe. You can't do that by saying that rationality is impossible unless such-and-such. Maybe there just isn't any coherent notion of rationality in the Everett interpretation?

Author: You misunderstand. I'm just saying that rationality requires diachronic consistency: that any rational strategy is a diachronically consistent strategy. So I'm constraining the space of rationally possible behaviours. If it turns out to be empty, of course, we're in trouble. But it won't: the Born-rule strategy is diachronically consistent and satisfies all the other axioms. All I'm doing is restricting (eventually to zero) the set of non-Born strategies.

Skeptic: What if the Born rule is also irrational?

Author: Which is to say: what if it violates some rationally required constraint on action? Then we're sunk. But it doesn't.

Skeptic: What about —

Author: Yes, yes, "it's rationally required to weight each branch equally". We'll come to that.

Incidentally, the very idea of composing acts to make further acts also presupposes diachronic consistency: only if an agent can think of future decisions he will make as *his decisions*, so that he can meaningfully make those decisions (for all that there is always some possibility that he will change his mind) does it make sense to consider composite acts.

The remaining rationality axioms are more specific to the Everettian context. Their precise statements get a bit more technical, so I phrase them fairly loosely here; as always, see section 7 for details and for reassurance there there isn't sleight-of-hand going on.

[6] For arguments that ascriptions of irrationality *only* make sense against a presumed backdrop of rationality, see Davidson (1973, 2004), Dennett (1987, pp. 83–116), and Lewis (1974).

[7] See the first part of Greg Egan's novel *Permutation City* for a science-fictional exploration of the idea — but notice that its plausibility relies on the copy's actions being causally relevant to the original, something not possible in the Everettian universe.

Microstate Indifference: An agent doesn't care what the microstate is provided it's within a particular macrostate.

Branching Indifference: An agent doesn't care about branching *per se*: if a certain measurement leaves his future selves in N different macrostates but doesn't change any of their rewards, he is indifferent as to whether or not the measurement is performed.

State Supervenience: An agent's preferences between acts depend only on what physical state they actually leave his branch in: that is, if $U\psi = U'\psi'$ and $V\psi = V'\psi'$, then an agent whose prefers U to V given that the initial state is ψ should also prefer U' to V' given that the initial state is ψ' — $U \succ^{\psi} V$ iff $U' \succ^{\psi'} V'$.

Solution Continuity: If for some state ψ $\widehat{U} \succ_{\psi} \widehat{U}'$, then sufficiently small permutations of $\widehat{U}$ and $\widehat{U}'$ will not change this.

Macrostate indifference is hopefully uncontroversial: it's built into the definition of macrostates, in fact. (The point being that an agent can have no practical control as to what state he gets, within a particular macrostate, on familiar statistical-mechanics and decoherence grounds.)

Solution continuity and branching indifference — and indeed problem continuity — can be understood in the same way, in terms of the limitations of any physically realisable agent. Any discontinuous preference order would require an agent to make arbitrarily precise distinctions between different acts, something which is not physically possible. Any preference order which could not be extended to allow for arbitrarily small changes in the acts being considered would have the same requirement. And a preference order which is not indifferent to branching *per se* would in practice be impossible to act on: branching is uncontrollable and ever-present in an Everettian universe.[8]

Skeptic: Why assume *a priori* that the rational strategy must be physically possible? Even if there is some strategy in an Everettian universe which counts as rational, maybe it's not physically possible to carry out that strategy.

Author: That's confused. Firstly, we already know there's at least one possible rational strategy: the Born rule. Secondly, what would it even be for a strategy to be rational, but physically impossible? By that token, the rational strategy for a trader is "always buy shares that are going to increase in value".

To be fair, a strategy might be literally impossible but be an idealization of a possible strategy — after all, perfect rationality itself is an idealization. One might *possibly* relax the assumption of Continuity on these

[8] The main source of branching is probably classically chaotic systems; see Zurek and Paz (1994) for technical details, and Wallace (2001) for discussion.

grounds (and I'll make some comments on that later), though I don't really think it's justified. But no strategy can approximate caring about branch number, as we'll see.

The other way to understand these assumptions is as prohibitions on strategies that just exploit artefacts of our model. The branching structure — including the well-defined number of branches associated with any act — is derived from the set of macrostates, which is in turn derived from decoherence. But as I argued in my earlier chapter in this volume that this structure has a significant degree of arbitrariness associated with it, primarily in terms of the coarseness of the grain of the macrostates (see also James Hartle's contribution). Put simply, in the actual physics there is no such thing as a well-defined branch number. Similarly, in the actual physics there is no division of the dynamics into discrete branching events followed by evolution of individual branches: branching, rather, is continuous. But if branching is always going on, and cannot be quantified in a non-arbitrary manner, then no strategy can be formulated which is other than indifferent to the presence of branching.

A quick defence of state supervenience would be: the agent's preferences supervene on the actual state of the branch; transformations which differ only in how they would affect non-actual quantum states do not differ in any relevant respect.

Skeptic: Hang on. This brings out a tacit assumption in the formalism you've adopted: the idea that acts can be represented by *single* unitary transformations rather than by *sequences* of unitary transformations. Why regard a sequence of measurements as decision-theoretically equivalent to a single measurement just because the same unitary transformation is enacted by both?

Author: Here's one possible defence. The agent is playing a sequence of games which result in rewards that he spends only after the sequence is done. In this case, what does he care about what happens during the brief period in which the games are being played (when having or not having rewards makes no difference to his status) — should he not care only about the state of the universe after the payouts are all made?

Skeptic: Well, that sounds intuitive, but so what? We're discussing *the Everett interpretation* — appeals to intuition are going to ring a little hollow here.

Author: Fair enough. A far better defence is to observe that caring about the final state only is the diachronic equivalent of branch indifference, and can be defended in the same way. There is no "real" branching structure beyond a certain fineness of grain, so the details of that structure can only be included in terms of their coarse-grained consequences.

Put another way: we could have defined our decision theory in terms of preferences, not over final states, but over consistent history spaces. But if we had done so, we would have needed both synchronic and diachronic indifference assumptions: indifference both to the fineness of grain of the history projectors at each time, and to the size of the temporal gaps between history projectors. Translated back into our setting, where we consider sequences of decisions made only over very short periods of time, the former assumption entails branch indifference and the latter entails that acts can be represented by single unitary transformations.

6 A quantum representation theorem

We can now prove, in succession, three results, the first three of which are (trivially) entailed by the fourth.

Equivalence lemma: If two acts assign the same weight to each reward, the agent must be indifferent between them.

Nullity lemma: An agent is indifferent to a possible outcome of an act iff that act has weight zero.

Dominance lemma: Suppose that two acts each only have two possible rewards r_1, r_2 as outcomes, with $r_1 \succ r_2$[9] and that the first act assigns a higher weight to r_1 than the second act does. Then the first act must be preferred to the second.

Born rule theorem: There is a utility function on the set of rewards, unique up to affine transformations, such that one act is preferred to another iff its expected utility, calculated with respect to this utility function and to the quantum-mechanical weights of each reward, is higher.

Since all these results are proved *formally* in section 8, my purpose in this section is explanation and not persuasion: I wish simply to show the general shape of the proof.

The equivalence lemma is best illustrated by examples (here I basically follow the argument of Wallace (2007)). For a simple case, suppose we have two acts (A and B, say): in each, a system is prepared in a linear superposition $\alpha\left|+\right\rangle + \beta\left|-\right\rangle$ and then measured in the $\{\left|+\right\rangle, \left|-\right\rangle\}$. On act A, a reward is then given if the result is '+'; on B, the same reward is given on '-' instead. The resultant states are

$$\text{A}: \;\; \alpha\left|+\right\rangle\otimes\left|\text{reward}\right\rangle + \beta\left|-\right\rangle\otimes\left|\text{no reward}\right\rangle; \tag{5}$$

$$\text{B}: \;\; \alpha\left|+\right\rangle\otimes\left|\text{no reward}\right\rangle + \beta\left|-\right\rangle\otimes\left|\text{reward}\right\rangle. \tag{6}$$

By erasure, there will exist acts available to the agent's future self in the reward branch (for both A and B) which erase the result of what was measured, leaving

[9] That is: with an act which returns some microstate in r_1 with certainty preferred to one which returns some microstate in r_2 with certainty; that this determines a well-defined ordering over rewards follows from microstate indifference.

only the reward. Performing these transformations, and the equivalent erasures in the no-reward branch, leaves

$$\text{A-plus-erasure:}\;\; \alpha\,|0\rangle\otimes|\text{reward}\rangle + \beta\,|0'\rangle\otimes|\text{no reward}\rangle\,; \tag{7}$$

$$\text{B-plus-erasure:}\;\; \beta\,|0\rangle\otimes|\text{reward}\rangle + \alpha\,|0'\rangle\otimes|\text{no reward}\rangle\,. \tag{8}$$

Now, by branch indifference, the agent's future selves are indifferent to whether this erasure is or is not performed. (Branch indifference is needed because we have no guarantee that erasures are non-branching; if we did, microstate indifference would suffice). So by diachronic consistency, the original agent is indifferent between A and A-plus-erasure, and between B and B-plus-erasure.

But now: if $\alpha = \beta$, then A-plus-erasure and B-plus-erasure leave the system in the same quantum state. So by state supervenience, the agent is indifferent between them. Since we know from ordering that preferences are transitive, the agent must also be indifferent between A and B. Indeed, we actually require only that $|\alpha| = |\beta|$, for phase differences too can be erased.

For a slightly more complicated case, suppose game C involves a 2-state system being prepared in state

$$\sqrt{2/3}\,|+\rangle + \sqrt{1/3}\,|-\rangle$$

and a reward being given on '+', and game D involves a 3-state system being prepared in state

$$\sqrt{1/3}(|+\rangle + |0\rangle + |-\rangle)$$

and a reward being given on '+' and on '0'. The resultant states are then

$$C:\; \sqrt{2/3}\,|+\rangle\otimes|\text{reward}\rangle + \sqrt{1/3}\,|-\rangle\otimes|\text{no reward}\rangle\,; \tag{9}$$

$$D:\; \sqrt{1/3}\,|+\rangle\otimes|\text{reward}\rangle + \sqrt{1/3}\,|0\rangle\otimes|\text{reward}\rangle + \sqrt{1/3}\,|-\rangle\otimes|\text{no reward}\rangle\,. \tag{10}$$

But by erasure, there is an act available for the future self of the agent in the 'reward' branch of game C which creates two equally-weighted branches:

$$|+\rangle\otimes|\text{reward}\rangle \longrightarrow \sqrt{1/2}\,|X\rangle\otimes|\text{reward}\rangle + \sqrt{1/2}\,|Y\rangle\otimes|\text{reward}\rangle \tag{11}$$

Since by branch indifference the agent's future self is indifferent to performing this act or not, by diachronic consistency the original agent is indifferent between C and C-plus-branching. But the state produced by C-plus-branching is

$$\text{C-plus-branching}:\; \sqrt{1/3}\,|X\rangle\otimes|\text{reward}\rangle + \sqrt{1/3}\,|Y\rangle\otimes|\text{reward}\rangle + \sqrt{1/3}\,|-\rangle\otimes|\text{no reward}\rangle\,. \tag{12}$$

By a generalisation of our earlier argument, the agent is indifferent between C-plus-branching and D, and so between C and D.

By arguments of this kind, the equivalence lemma can be proved for any act with finitely many outcomes. The null and dominance lemmas are easy further steps, using the second clause of diachronic consistency.

We are now nearly done: the remainder of the proof is actually a standard decision-theoretic method for constructing utilities. Pick two rewards R and S

with $R \succ S$, and assign R utility 1 and S utility 0. For any reward T satisfying $R \succeq T \succeq S$, there is a unique number $U(T)$ such that the agent is indifferent between getting T with certainty, and getting R on a branch of weight $U(T)$ and S otherwise. (We need continuity to establish this and rule out the possibility of rewards whose utilities differ only infinitesimally.)

Now consider an act which leads to rewards R, S, T with weights w(R), w(S) and w(T) respectively. The agent's future selves in the T branch are indifferent between doing nothing and performing an act that delivers R with weight $U(T)$ and S otherwise. Applying diachronic consistency once more, the original agent is indifferent between the original act and an act which delivers R with weight $w(R) + w(T)U(T)$ and an act which delivers S with weight $w(S) + (1 - U(T))w(T)$. Note that the utilities of these acts are the same: in this particular case, the agent is indifferent between two acts iff they have the same utility. Generalising the argument, and applying the dominance lemma, tells us that one act is preferred to another iff its utility is higher.

The continuity axioms play only a limited role in these arguments. They serve to rule out situations where two rewards are infinitesimally, or infinitely, different in value; they are also required to handle the generalisation to acts which have infinitely many rewards as possible outcomes.

7 Formal statement of the axioms

As promised, in this section and the next I lay out the formal version of my decision theory and its associated proofs. The reader who is happy to take on trust my mathematics — and my reassurances that there has been no sleight of hand — is welcome to skip to section 9.

A *quantum decision problem* is specified by:

- A separable Hilbert space $\mathcal{H}$. Given a set $\mathcal{S}$ of subspaces of $\mathcal{H}$, I write $\vee\mathcal{S}$ (the *disjunction* of $\mathcal{S}$) for the closure of the span of $\cup\mathcal{S}$, and $\wedge\mathcal{S}$ (the *conjunction* of $\mathcal{S}$) for the closure of $\cap\mathcal{S}$; Given subspaces E and F, I define $E \vee F = \vee\{E, F\}$ and likewise for $\wedge$, and I write Π_E for the projector onto E.

- A complete Boolean algebra $\mathcal{E}$ of subspaces of $\mathcal{H}$, the *event space*. (So $\mathcal{E}$ contains $\mathcal{H}$ and is closed under $\vee$, $\wedge$, and taking the complement.) I define a *partition* of an event E to be a set of mutually orthogonal events whose conjunction is E.

- A subset $\mathcal{M}$ of $\mathcal{E}$, the *macrostates*, such that for any event E, there is a partition of E by macrostates.

- For each $E \in \mathcal{E}$, a set $\mathcal{U}_E$ of unitary operators from E into $\mathcal{H}$, which we call the set of *acts available at E*. We write $\mathcal{O}_U$ for the smallest event containing the range of the act U[10] and require that the choice of available

[10]We can define $\mathcal{O}_U$ explicitly as the conjunction of all events containing the range of U; this suffices to show that $\mathcal{O}_U$ is well-defined.

acts satisfies:

1. *Restriction*: If $E, F \in \mathcal{E}$ and $F \subset E$, then if U is available at E then the unitary map $U|_F$, defined by $U\psi = U|_F\psi$ whenever $\psi \in F$, is available at F.
2. *Composition*: If U is available at E, and V is available at $\mathcal{O}_U$, then VU is available at E.
3. *Indolence*: For any event E, if there are any acts available at E then the identity $\widehat{1}_E$ is available at E. (More precisely, the embedding map of E into $\mathcal{H}$ is available at E.)
4. *Continuation*: If U is available at some E, then there is some act available at $\mathcal{O}_U$.
5. *Irreversibility*: If U is available at $E \vee F$, $\mathcal{O}_{U|_E} \wedge \mathcal{O}_{U|_F} = \emptyset$.

- A partition $\mathcal{R}$ of $\mathcal{E}$ (that is, a set of mutually orthogonal elements of $\mathcal{E}$ whose disjunction is $\mathcal{H}$), the set of *rewards.* These represent payoffs an agent could get.

The simplest choice of macrostates and event space is to pick some particular set of orthogonal subspaces of $\mathcal{H}$ whose disjunction is $\mathcal{H}$, take this as $\mathcal{M}$, and take $\mathcal{E}$ to be the set of all disjunctions of subsets of $\mathcal{M}$; this is the sense of "macrostate" and "event" used in the informal version of the proof. However, we could equally well take $\mathcal{E}$ to be an arbitrary Boolean algebra of subspaces and define $\mathcal{E} = \mathcal{M}$. (As was noted previously, this sort of formalisation might be more appropriate for decision problems with a less natural discrete structure.)

Rays within $\mathcal{H}$, as usual, are called states. I adopt the usual convention of representing a ray by any vector within it and of blurring the distinction between the two; I do not require that vectors representing states be normalised. (This is just for notational convenience.) If $\mathcal{B}(E, \mathcal{H})$ is the set of unitary maps from E into $\mathcal{H}$, it can naturally be regarded as a subset of $\mathcal{B}(\mathcal{H}, \mathcal{H})$ by identifying U with $U\Pi_E$; as such, $\mathcal{B}(E, \mathcal{H})$ inherits the norm topology.

I introduce a few derived concepts. The *weight* $\mathcal{W}_\psi(E|U)$ of an event E with respect to a state ψ and an act U is defined by

$$\mathcal{W}_\psi(E|U) = \|\Pi_E U \left|\psi\right\rangle\|^2 = \left\langle\psi\right| U^\dagger \Pi_E U \left|\psi\right\rangle . \tag{13}$$

A *reward function* is any function from $\mathcal{R}$ to $[0, 1]$ such that $\sum_{r\in\mathcal{R}} w(r) = 1$. Any pair of a state $\psi \in E$ and an act U available at E determines a reward function

$$R_{\psi,U}(r) = \mathcal{W}_\psi(r|U) \tag{14}$$

which I call the *characteristic reward function* of U and ψ.

A set $\mathcal{F}$ of events is *available* if they are mutually orthogonal and there is at least one act available at $\vee\mathcal{F}$. (An event is available iff its singleton set is available).

Finally, if $\mathcal{S}$ is any set of rewards, I say that an act A *has rewards in* $\mathcal{S}$ iff its range is a subset of $\vee\mathcal{S}$. If u is a real function of $\mathcal{S}$ and U is an act whose rewards are in $\mathcal{S}$, the *expected utility* of U with respect to a state ψ (and, tacitly, with respect to u) is

$$\mathrm{EU}_\psi(U) = \sum_{r\in\mathcal{S}} \mathcal{W}_\psi(r|U)u(r) \equiv \sum_{r\in\mathcal{S}} R_{\psi,U}(r)u(r). \tag{15}$$

Stating the richness axioms is a little fiddly, because of the need to make sure not only that certain acts (erasures, branchings etc) are available everywhere, but to make sure that they are available on multiple branches concurrently. To state them in a concise way, I make the following definitions. Firstly, if $\mathcal{P} = \{p_1, p_2, \ldots\}$ is a (countable or finite) set of positive real numbers whose sum is unity, and $\psi \in M \subset r$ for some state ψ, macrostate M, and reward r, then a $\mathcal{P}$-branching of ψ is some act U available at M such that $\mathcal{O}_U \subset r$ and such that there is a partition $\mathcal{M} = \{M_1, M_2, \ldots\}$ of $\mathcal{O}_U$ by macrostates with $\mathcal{W}_\psi(M_i|U) = p_i$. (Informally, a $\mathcal{P}$-branching is an act which splits the agents branch into many branches, each having the same weight as an element of $\mathcal{P}$, but without changing the rewards that the agent gets.)

Secondly, if M and M' are macrostates with $M \subset r$ and $M' \subset r$ for some reward r, and ψ, ψ' are states in M, M' respectively, then an *erasure* of ψ and ψ' is a pair of acts U, U' available at M and M' respectively, such that $\mathcal{O}_U$ and $\mathcal{O}_{U'}$ are both subsets of r and $U\psi = U'\psi'$.

And thirdly, if $\mathcal{F}$ is an available set of events, an *act function* $\mathcal{U}$ for that set is a function which assigns to each $F \in \mathcal{F}$ an act $\mathcal{U}(F)$ available at F. An act function is *compatible* if

$$\sum_{F\in\mathcal{F}} \mathcal{U}(F)\Pi_F \tag{16}$$

is available at $\vee\mathcal{F}$.

The richness axioms are now stateable:

Reward availability: Suppose that $\mathcal{F}$ is an available set of macrostates and f is a function from $\mathcal{F}$ into rewards.

Then there is a compatible act function $\mathcal{U}$ for $\mathcal{F}$ with $\mathcal{U}(F) \subset f(F)$ for all $F \in \mathcal{F}$.

Branching availability: Suppose that $\mathcal{F}$ is an available set of macrostates and for each $F \in \mathcal{F}$, ψ_F is a nonzero state in F and $\mathcal{P}_F$ is a (finite or countable) set of positive real numbers summing to unity.

Then there is a compatible act function $\mathcal{U}$ for $\mathcal{F}$ such that for each $F \in \mathcal{F}$, $\mathcal{U}(F)$ is a $\mathcal{P}_F$-branching of ψ_F.

Erasure: Suppose that $\{r_1, r_2, \ldots\}$ is a (finite or countable) set of rewards, that $\mathcal{M} = \{M_1, M_2, \ldots\}$ and $\mathcal{N} = \{N_1, N_2, \ldots\}$ are two available sets of macrostates with $M_i \subset r_i$ and $N_i \subset r_i$, and that for each i, $\psi_i \in M_i$ and $\varphi_i \in N_i$ are nonzero states.

Then there are compatible act functions $\mathcal{U}$ for $\mathcal{M}$ and $\mathcal{V}$ for $\mathcal{N}$ such that, for each i, $(\mathcal{U}(M_i), \mathcal{V}(N_i))$ is an erasure of ψ_i and φ_i.

Problem Continuity: For every available E, the set of acts available at E is an open subset (in operator norm topology) of the set of unitary maps from E to $\mathcal{H}$.[11]

Notice that reward availability and preparation together entail that for any reward function and any $\psi \in E$, there is an act U available at E such that ψ and U have that reward function as their characteristic reward function.

We now define a *state-dependent solution* to a decision problem as specified by an assigment to every available macrostate E, and every state $\psi \in E$, of a two-place relation $\succ^{\psi}$ on the acts available at E. (Strictly our notation should include E but for simplicity, its value will always be tacit.)

We call an event N *null* for a given state ψ and act U iff, whenever acts V_1 and V_2 are identical on the complement of N, $V_1 U \sim^{\psi} V_2 U$. (So an event is null if the agent doesn't care what happens to his future selves, if any, in the branch defined by that event. We will shortly see that, as expected, an event is null iff there are in fact no such future selves.) It is easy to see that any finite union of null sets is null, as is any subset of a null set.

We can now state the rationality axioms:

Ordering: For every ψ for which it is defined, $\succ^{\psi}$ is a total ordering. That is: it is transitive, asymmetric, and the relation $\sim^{\psi}$, defined by $E \sim^{\psi} F$ iff neither $E \succ^{\psi} F$ nor $F \succ^{\psi} E$, is an equivalence relation. (As usual, we write '$E \succeq^{\psi} F$' as an abbreviation for 'either $E \succ^{\psi} F$ or $E \sim^{\psi} F$'.)

Diachronic Consistency: Suppose U is available at E, and V_1 and V_2 are available at $\mathcal{O}_U$. Then:

(i) If there is some partition $\mathcal{P}$ of $\mathcal{O}_U$ into macrostates such that $V_1|_E \succeq^{\Pi_E U \psi} V_2|_E$ for every element E of the partition not null with respect to ψ and U, then $V_1 U \succeq^{\psi} V_2 U$.

(ii) If in addition, $V_1|_E \succ^{\Pi_E U \psi} V_2|_E$ for at least one such E, then $V_1 U \succ^{\psi} V_2 U$.

Macrostate indifference: If:

- U, V are acts available at M;
- U', V' are acts available at M';
- $\mathcal{O}_U \subset M_1 \wedge r_1$ and $\mathcal{O}_{U'} \subset M_1 \wedge r_2$ for some macrostate M_1 and reward r_1;
- $\mathcal{O}_V \subset M_2 \wedge r_2$ and $\mathcal{O}_{V'} \subset M_2 \wedge r_2$ for some macrostate M_2 and reward r_2

then for any ψ, ψ' with $\psi \in M$ and $\psi' \in M'$, $U \succeq^{\psi} V$ iff $U' \succeq^{\psi'} V'$.

[11] The operator norm topology on the set of linear maps between normed spaces V and W is defined by the norm $\|U\| = \sup\{\|Ux\| : \|x\| = 1\}$. The set of unitary maps from E to $\mathcal{H}$ is a subset of the set of all maps between those two spaces, and inherits the latter's topology.

Branching indifference: If:

- r is a reward;
- M is a macrostate with $M \subset r$;
- U is available at M;
- $\psi \in M$ and $U\psi \in r$

then $U \sim^{\psi} \widehat{1}_M$.

State supervenience: If:

- $\psi \in E$ and $\psi' \in E'$ for macrostates E, E';
- U and V are available at E, and U' and V' are available at E';
- $U\psi = U'\psi'$ and $V\psi = V'\psi'$

then $U \succ^{\psi} V$ iff $U' \succ^{\psi'} V'$.

Solution continuity: If E is a macrostate and $\psi \in E$, if U, U' are available at E, and if $U \succ^{\psi} U'$, then in the space of unitary maps from E into $\mathcal{H}$ there are neighbourhoods (in norm topology) $\mathcal{N}, \mathcal{N}'$ of U, U' respectively such that any act in $\mathcal{N}$ available at E is preferred (at ψ) to any act in $\mathcal{N}'$ available at E.

Given a solution to a quantum decision problem, we can use it to define a preference ordering on rewards: for any two rewards, $r_1 \succ r_2$ iff there is some macrostate E, some state $\psi \in E$, and acts U_1, U_2 available at E such that $\mathcal{O}_{U_i} \subset r_i$ and $U_1 \succ^{\psi} U_2$. Provided that the problem is reward-available and the solution is macrostate-indifferent and branching-indifferent, this preference order is a total ordering on $\mathcal{R}$. If r and s are rewards with $r \preceq s$, I will say that a reward t is between r and s iff $s \succeq t \succeq t$; I write $[r, s]$ for the set of rewards between r and s.

If $\mathcal{M}$ consists of some set of orthonormal subspaces (as in the informal proof), then this observation more or less exhausts the usefulness of macrostate indifference. At the other extreme, if $\mathcal{M}$=$\mathcal{E}$ then macrostate indifference actually entails branch indifference. The distinction between the axioms, then, is a matter of how we mathematically represent the branching structure — which is appropriate, since the motivation for branching indifference itself is that the details of that structure are an unphysical artefact of the mathematics.

(The mathematically inclined reader may be wondering at this point if the axioms are consistent. To show that they are, consider the following model. Let $\mathcal{H}_R$ be a two-dimensional Hilbert space with an orthogonal basis $\{|+\rangle, |-\rangle\}$; for each $N > 0$ let $\{\mathcal{H}_N\}$ be an N-dimensional Hilbert space with an orthonormal basis $\{|N,1\rangle, |N,2\rangle, \ldots |N,N\rangle\}$.

Now: take the Hilbert space of our decision problem to be

$$\mathcal{H} = \mathcal{H}_R \otimes (\oplus_{I=1}^{\infty} \mathcal{H}_I), \tag{17}$$

so that a complete basis of states is

$$|\pm\rangle \otimes |N, M\rangle \quad (M \leq N), \tag{18}$$

and take the macrostates to consist of all the one-dimensional subspaces spanned by each of these states, and the events to be all disjunctions of macrostates. The available events are all those which are contained in some fixed $\mathcal{H}_R \otimes \mathcal{H}_N$, and the acts available at an available event contained in $\mathcal{H}_R \otimes \mathcal{H}_N$ are all unitary maps from $\mathcal{H}_R \otimes \mathcal{H}_N$ to $\mathcal{H}_R \otimes \mathcal{H}_{N'}$, with $N' > N$. The reward subspaces are $\mathcal{H}^{\pm} = \{\text{Span}\,|\pm\rangle\} \otimes \mathcal{H}$. Finally, an act U is preferred to an act U' at $|\psi\rangle$ iff

$$\left\| (|+\rangle \langle +| \otimes \widehat{1}) U |\psi\rangle \right\| > \left\| (|+\rangle \langle +| \otimes \widehat{1}) U' |\psi\rangle \right\|. \tag{19}$$

I leave readers to satisfy themselves that this system does indeed obey the axioms; the preference order is, of course, the Born rule.)

8 Formal statement and proof of the representation theorem

Equivalence Lemma: Suppose that:

(i) $\mathcal{P}$ is a quantum decision problem satisfying erasure, branch availability and reward availability;

(ii) $\succ^{\psi}$ is a state-dependent solution to $\mathcal{P}$ satisfying ordering, diachronic consistency, macrostate indifference, branching indifference, and state supervenience;

(iii) U and V are available at E, and U' and V' are available at E';

(iv) $\psi \in E$ and $\psi' \in E'$;

(v) $R_{\psi,U} = R_{\psi',U'}$ and $R_{\psi,V} = R_{\psi',V'}$.

(vi) The reward functions of the acts are each non-zero for only finitely many rewards.

Then $U \succ^{\psi} V$ iff $U' \succ^{\psi'} V'$.

Proof: For each reward r for which $R_{\psi,U}(r) \neq 0$, let $\mathcal{M}_r$ and $\mathcal{N}_r$ be partitions of $\mathcal{O}_U \wedge r$ and $\mathcal{O}_{U'} \wedge r$ respectively, and let $\#M_r$ and $\#N_r$ be the number of elements (finite or infinite) in $\mathcal{M}_r$ and $\mathcal{N}_r$ respectively.

Define the sets $\mathcal{P}_r$ (for each r)

$$\mathcal{P}_r = \{\mathcal{W}_{\psi'}(N|U')/\mathcal{W}_{\psi'}(r|U') : N \in \mathcal{N}_r\} \tag{20}$$

These are sets of positive real numbers summing to unity, so by branching availability there is an act W available at $\mathcal{O}_U$ such that, for each r and each $M \in \mathcal{M}_r$, $W|_M$ is a $\mathcal{P}_r$-branching of $\Pi_M U\psi$: it splits $\Pi_M U\psi$, which has weight $\mathcal{W}_\psi(M|U)$, into $\#\mathcal{N}_r$ states, one for each $N \in \mathcal{N}_r$, with weights $\mathcal{W}_\psi(M|U) \times$

$\mathcal{W}_{\psi'}(N|U')/\mathcal{W}_{\psi'}(r|U')$. There is therefore[12] a partition $\mathcal{W}$ of $\mathcal{O}_W$ into macrostates, such that:

- For each reward r there are $\#\mathcal{M}_r \times \#\mathcal{N}_r$ elements of $\mathcal{W}$ in r.
- Each such element can be labelled by pairs of elements from $\mathcal{M}_r$ and $\mathcal{N}_r$: let us write it as $K^r_{M,N}$.
- $\mathcal{W}_\psi(K^r_{M,N}|WU) = \mathcal{W}_\psi(M|U) \times \mathcal{W}_{\psi'}(N|U')/\mathcal{W}_{\psi'}(r|U')$.

Furthermore, by branching indifference, $W|_M \sim^{\Pi_M U\psi} \widehat{1}_M$ for any macrostate M, and hence by diachronic consistency, $WU \sim^\psi U$.

Applying the same procedure with U and U' reversed yields an act W' such that $W'U \sim^\psi U$, and a partition $\mathcal{W}'$ of $\mathcal{O}_{W'}$ by macrostates, such that

- For each reward r there are $\#\mathcal{M}_r \times \#\mathcal{N}_r$ elements of $\mathcal{W}'$ in r.
- Each such element can be labelled by pairs of elements from $\mathcal{M}_r$ and $\mathcal{N}_r$: we write it as $K'^r_{M,N}$.
- $\mathcal{W}_{\psi'}(K'^r_{M,N}|WU) = \mathcal{W}_\psi(M|U) \times \mathcal{W}_{\psi'}(N|U')/\mathcal{W}_\psi(r|U)$.

But since

$$\mathcal{W}_\psi(r|U) \equiv \mathcal{R}_{\psi,U}(r) = \mathcal{R}_{\psi',U'}(r) \equiv \mathcal{W}_{\psi'}(r|U'), \tag{21}$$

it follows that $\mathcal{W}_\psi(K^r_{M,N}|WU) = \mathcal{W}_{\psi'}(K'^r_{M,N}|W'U')$.

So we have constructed acts W, W' and partitions $\mathcal{W} = \{W_1, \ldots\}$, $\mathcal{W}' = \{W'_1, \ldots\}$ of $\mathcal{O}_W$, $\mathcal{O}_{W'}$ by macrostates such that:

1. For any i, W_i exists iff W'_i does (i. e. , the two partitions have the same number of elements) and there is some reward r such that W_i and W'_i are elements of r.
2. $\mathcal{W}_\psi(W_i|WU) = \mathcal{W}_\psi(W'_i|W'U)$ for all W_i.

Now define

$$\chi_i = \Pi_{W_i} WU\psi / \|\Pi_{W_i} WU\psi\| \tag{22}$$

and

$$\chi'_i = \Pi_{W'_i} W'U'\psi' / \|\Pi_{W'_i} W'U'\psi'\|. \tag{23}$$

By erasure, there exist acts X, X' available at $\mathcal{O}_W$, $\mathcal{O}_{W'}$ such that $(X|_{W_i})\chi_i = (X'|_{W'_i})\chi'_i$. By branching indifference, $X|_{W_i} \sim^{\chi_i} \widehat{1}_{W_i}$, so by diachronic consistency, $XWU \sim^\psi WU \sim^\psi U$; similarly, $X'W'U' \sim^{\psi'} U'$.

Since

$$XWU\psi = \sum_i \mathcal{W}_\psi(W_i|WU)(X|_{W_i})\chi_i, \tag{24}$$

it follows that $XWU\psi = X'W'U'\psi'$.

[12]We appeal here to the irreversibility requirement on decision problems.

So: for U and U', we have found acts $Y = XWU$ and $Y' = X'W'U'$ such that $U \sim^{\psi} Y$, $U' \sim^{\psi'} Y'$, and $Y\psi = Y'\psi'$. Repeating this process for V and V', we can find acts Z, Z' such that $Z \sim^{\psi} V$, $Z' \sim^{\psi'} V'$, and $Z\psi = Z'\psi'$. The conclusion now follows immediately from state supervenience. $\square$

Because of the equivalence lemma, there is a unique total ordering defined on the set of all reward functions, which we once again write as $\succ$ (note that it is state-independent).

Nullity Lemma: Suppose that:

(i) $\mathcal{P}$ is a quantum decision problem satisfying erasure, branch availability and reward availability;

(ii) $\succ^{\psi}$ is a state-dependent solution to $\mathcal{P}$ satisfying ordering, diachronic consistency, macrostate indifference, branching indifference, and state supervenience;

(iii) There exist rewards r, s with $r \succ s$.

Then an event E is null with respect to a state ψ and an act U iff $\langle\psi| U^{\dagger}\Pi_E U |\psi\rangle = 0$.

Proof: Let $\langle\psi| U^{\dagger}\Pi_E U |\psi\rangle = \alpha$. An event is null if and only if, given acts V and W available at $\mathcal{O}_U$ which are identical except on E, $VU \sim^{\psi} WU$. Given the Equivalence Lemma, any two such acts are equivalent whenever they have the same weight function, so if E is null for ψ and U, any event E' is null with respect to some U' and ψ' whenever $\langle\psi'| U'^{\dagger}\Pi'_E U' |\psi'\rangle = \alpha$. If $\alpha > 0$, then $\alpha > 1/N$ for some N. By combining branch availability with reward availability, we can construct some act V and state φ with weight function

$$\begin{aligned}
\mathcal{W}_{\varphi}(E_1|V) &= 1/N \\
\mathcal{W}_{\varphi}(E_2|V) &= \alpha - 1/N \\
\mathcal{W}_{\varphi}(E_3|V) &= 1 - \alpha
\end{aligned}$$

$E_1 \vee E_2$ is null (wrt φ and V), hence E_1 is, hence any event with weight $1/N$ is. Applying branch availability and reward availability again, we can find φ', W and $F_1, \ldots F_N$ such that $\mathcal{W}_{\varphi'}(F_i|W) = 1/N$. Each F_i is null wrt φ' and W, hence so is $\mathcal{E}$. This contradicts premise (iii), since if all events are null then all rewards are equivalent.

Conversely, suppose that some event has weight zero. Its nullity now follows from state supervenience, since no change to the physical state is enacted by any transformation restricted to that event. $\square$

Dominance Lemma: Suppose that

(i) $\mathcal{P}$ is a quantum decision problem satisfying erasure, branch availability and reward availability;

(ii) $\succ^\psi$ is a state-dependent solution to $\mathcal{P}$ satisfying ordering, diachronic consistency, macrostate indifference, branching indifference, and state supervenience;

(iii) s, t are rewards with $s \succ t$.

(iv) $f[\alpha]$ is the reward function defined by $f[\alpha](s) = \alpha$, $f[\alpha](t) = 1 - \alpha$, $f[\alpha](r) = 0$ for all other r.

Then $f[\alpha] \succ f[\beta]$ iff $\alpha > \beta$.

Proof: This is an easy corollary of the Nullity Lemma. Suppose $\alpha > \beta$, then by branch availability and reward availability, there will be some act A and state φ with weight function

$$\begin{aligned} \mathcal{W}_\varphi(E_1|A) &= \beta \\ \mathcal{W}_\varphi(E_2|A) &= \alpha - \beta \\ \mathcal{W}_\varphi(E_3|A) &= 1 - \alpha \end{aligned}$$

By reward availability there exist sets of compatible acts $\{U_1, U_2, U_3\}$ and $\{V_1, V_2, V_3\}$ such that U_i and V_i are available at E_i, and such that U_1, V_1 and U_2 have outcomes all lying in s and V_2, U_3 and V_3 have outcomes all lying in t. By macrostate indifference and branching indifference $U_i \simeq^\chi V_i$ for any $\chi \in E_i$ and in particular $U_2 \succ^\chi V_2$ for any $\chi \in E_2$.

If we define

$$W_\alpha = U_1 \Pi_{E_1} + U_2 \Pi_{E_2} + U_3 \Pi_{E_3} \tag{25}$$

and

$$W_\beta = V_1 \Pi_{E_1} + V_2 \Pi_{E_2} + V_3 \Pi_{E_3} \tag{26}$$

then by diachronic consistency, since E_2 is not null then $W_\alpha \cdot A \succ^\psi W_\beta \cdot A$. But the reward functions of $W_\alpha \cdot A$ and $W_\beta \cdot A$ are $f[\alpha]$ and $f[\beta]$ respectively, and the conclusion follows.

Utility Lemma: Suppose that:

(i) $\mathcal{P}$ is a quantum decision problem satisfying erasure, branch availability, and reward availability;

(ii) $\succ^\psi$ is a state-dependent solution to $\mathcal{P}$ satisfying ordering, diachronic consistency, macrostate indifference, branching indifference, and state supervenience;

(iii) s, t are rewards with $s \succ t$.

(iv) u_s, u_t are real numbers with $u_s > u_t$.

Then there is a unique real function u on the set $[t, s]$ of rewards between t and s such that for any macrostate E, any state $\psi \in E$, and any two acts U, V available at E whose rewards lie a finite subset of $\mathcal{S}$,

$$U \succ_\psi V \text{ whenever } \mathrm{EU}_\psi(U) > \mathrm{EU}_\psi(V) \tag{27}$$

(where the expected utilities are defined with respect to u, of course) and such that $u(s) = u_s$ and $u(t) = u_t$.

Proof: For simplicity we assume $u_s = 1$ and $u_t = 0$ (other values lead to a simple affine transformation of the utility function). We define the following reward functions: $f[\alpha]$ is defined as in the Dominance Lemma, and $g[r]$ is defined by $g[r](r') = \delta_{r,r'}$.

We now define $u(r)$ by

$$u(r) = \mathrm{lub}\{\alpha : g[r] \succ f[\alpha]\}. \tag{28}$$

Let $\{u_n(r)\}$ be a sequence of functions such that $u_m(r) \leq u(r)$ and $\lim_{n\to\infty} u_n(r) = u(r)$, and let U be any act available at E whose rewards lie in $\mathcal{S}$. We write E_r for $\mathcal{O}_U \wedge r$ and χ_r for the normalised projection of ψ onto E_r.

From branching availability and reward availability, for each n we can find a compatible set of states $\{A_n(r) : R_{\psi,U}(r) \neq 0\}$ such that $A_n(r)$ is available at E_r and A_n has reward function $f[u_n(r)]$; we define $\mathcal{A}_n = \sum_{r\in\mathcal{S}} A_n(r)\Pi_{E_r}$. By construction, $\widehat{1}_{E_r} \succeq^{\chi_r} A_n(r)$ for all r and n, so by diachronic consistency $U \succeq^{\psi} \mathcal{A}_n \cdot U$.

By definition, the reward function of $\mathcal{A}_n \cdot U$ (with respect to ψ) is $f[\lambda_n]$, where

$$\lambda_n = \sum_{r\in\mathcal{S}} \mathcal{W}_\psi(r|U) u_n(r). \tag{29}$$

So if $f[U]$ is the reward function of U (with respect to ψ), we have established that $f[U] \succeq f[\lambda_n]$, and hence by the Dominance lemma, $f[U] \succeq f[\lambda]$ whenever $\lambda < \lambda_n$ for some n. Since $u_n(r) \to u(r)$ for each n and r, $\lambda_n \to \mathrm{EU}_\psi(U)$,and hence $f[U] \succ f[\lambda]$ whenever $\lambda < \mathrm{EU}_\psi(U)$. Applying the same argument with a decreasing sequence, $f[U] \prec f[\lambda]$ whenever $\lambda > \mathrm{EU}_\psi(U)$.

Now suppose that U and V are two such acts with $\mathrm{EU}_\psi(U) > \mathrm{EU}_\psi(V)$. Then for any α lying between the two expected utilities, there will exist an act W with reward function (wrt ψ) $f[\alpha]$. We have proved that $U \succ^\psi W$, and $W \succ^\psi V$, so it follows that $U \succ^\psi V$.

To see that this utility function is unique, note that if there were another utility function u' we could construct acts whose utilities were the same as calculated by this second utility, but not as calculated by the first; this contradicts the requirements on u'. $\Box$

Born Rule Theorem: Suppose that:

(i) $\mathcal{P}$ is a quantum decision problem satisfying erasure, branch availability, reward availability and problem continuity;

(ii) $\succ^\psi$ is a state-dependent solution to $\mathcal{P}$ satisfying ordering, diachronic consistency, macrostate indifference, branching indifference, state supervenience, and solution continuity.

Then there is a function u on the rewards of $\mathcal{P}$, unique up to positive affine transformations, such that if EU denotes the expected utility with respect to this function,

$$U \succ^{\psi} V \text{ iff } \mathrm{EU}_{\psi}(U) > \mathrm{EU}_{\psi}(V). \tag{30}$$

Proof: Note that problem continuity and solution continuity jointly entail that if $U \succ^{\psi} U'$, there are neighborhoods $\mathcal{N}, \mathcal{N}'$ of U and U' respectively such that all acts in $\mathcal{N}$ and $\mathcal{N}'$ are available and all acts in $\mathcal{N}$ are preferred (given ψ) to all acts in $\mathcal{N}'$. For simplicity I shall refer to this simply as continuity.

We begin by proving that if $s \succ r_1 \succeq r_2 \succ t$, then if the utilities determined by the Utility Lemma (via this choice of s and t) for r_1 and r_2 coincide, then $r_1 \sim r_2$. Let this utility function be u and again, for convenience take $u(s) = 1$ and $u(t)$=0. Fix E and $\psi \in E$, and let U_1 and U_2 be acts available at E whose ranges lie in r_1 and r_2 respectively (by reward availability, some such acts exist). If $r_1 \succ r_2$, then $U_1 \succ^{\psi} U_2$. By continuity, there must exist neighborhoods $\mathcal{N}_1$, $\mathcal{N}_2$ of U_1 and U_2 such that any available act in $\mathcal{N}_1$ is preferred (given ψ) to any available act in $\mathcal{N}_2$.

Now let $f_1[\alpha]$ and $f_2[\alpha]$ be reward functions with $f_1[\alpha](r_1) = 1-\alpha$, $f_1[\alpha](t) = \alpha$ and $f_2[\alpha](r_2) = 1-\alpha$, $f_2[\alpha](s) = \alpha$. By branch availability and reward availability, there must exist some α, and some acts $U_{i,\alpha}$, such that $U_{i,\alpha} \in \mathcal{N}_i$ and the reward function of $U_{i,\alpha}$ (with respect to ψ) is $f_i[\alpha]$.

So we have that $U_{1,\alpha} \succ U_{2,\alpha}$. But $\mathrm{EU}_{\psi}(U_{1,\alpha}) < \mathrm{EU}(U_1) \equiv u(r_1)$, and $\mathrm{EU}_{\psi}(U_{2,\alpha}) > \mathrm{EU}(U_2) \equiv u(r_2)$. So by the Utility lemma we must have that $u(r_1) > u(r_2)$.

We can now define a utility function for the whole of $\mathcal{R}$. For any rewards r_1, r_2 with $r_1 \succ r_2$, and any real numbers x_1, x_2 with $x_1 > x_2$, I will write $u[r_1, r_2, x_1, x_2]$ for the unique utility function determined on $[r_2, r_1]$ by setting the utility of r_i to x_i.

Now, let s, t be any two rewards with $s \succ t$ (if there are no such rewards, the theorem is true trivially). I define the utility of any reward r by:

- If $s \succeq r \succeq t$, $u(r) = u[s, t, 1, 0](r)$.
- If $r \succ s$, $u(r)$ is the unique value fixed by requiring that $u[r, t, u(r), 0](s) = 1$.
- If $t \succ r$, $u(r)$ is the unique value fixed by requiring that $u[s, r, 1, u(r)](s) = 0$.

(Notice that this definition relies on the assumption that the utilities of s and t are guaranteed to be distinct.)

I now prove that for acts with finitely many rewards, if $U_1 \succ^{\psi} U_2$ then $EU_{\psi}(U_1) > EU_{\psi}(U_2)$. For suppose that $U_1 \succ^{\psi} U_2$. By continuity, if f is the reward function of $\widehat{U}_1$ (with respect to ψ) then it will be possible to find some act V with reward function g such that, for some rewards r_1 and r_2 with $r_1 \succ r_2$:

- $V \succ^{\psi} U$;

- If $r \neq r_1$ and $r \neq r_2$, $g(r) = f(r)$;
- $g(r_1) < f(r_1)$; $g(r_2) > f(r_2)$.

This means that we must have $\mathrm{EU}_\psi(V) \geq \mathrm{EU}_\psi(U_2)$; since $\mathrm{EU}_\psi(V) < \mathrm{EU}_\psi(U_1)$, it follows that $EU_\psi(U_1) > EU_\psi(U_2)$.

This suffices to prove the Born Rule Theorem under the assumption that any act has only finitely many non-null rewards. To extend to the infinite case, let U_1 and U_2 be arbitrary acts, and suppose for some ψ that $U_1 \succ^\psi U_2$. By continuity, if f_1 and f_2 are the reward functions (given ψ) of U_1 and U_2, it will be possible to find a finite subset $\mathcal{R}_0$ of $\mathcal{R}$, and acts V_1, V_2 with reward functions g_1, g_2, such that:

- $V_1 \succ^\psi V_2$;
- $g_i(r) = f_i(r)$ for $r \in \mathcal{R}_0$;
- If $r \notin \mathcal{R}_0$, then $g_1(r) = s$, and $g_2(r) = t$, where $s \succ t$.

Since V_1 and V_2 have only finitely many non-null rewards, $\mathrm{EU}_\psi(V_1) > \mathrm{EU}_\psi(V_2)$. But by construction $\mathrm{EU}_\psi(U_1) > \mathrm{EU}_\psi(V_1)$ and $\mathrm{EU}_\psi(U_2) < \mathrm{EU}_\psi(V_2)$, so $\mathrm{EU}_\psi(U_1) > \mathrm{EU}_\psi(U_2)$. $\Box$

9 Other proposed strategies for action

In the nine years since Deutsch's original paper on decision-theoretic probability, a bewildering variety of alternative strategies for rational action have been proposed in the literature and in discussion. Some of these strategies have independent motivations; some are purely meant as counter-examples; all contradict the Born rule, and so all violate the decision-theoretic axioms of this paper.

This being the case, perhaps there is little need to discuss the alternative strategies: a proof is a proof. On the other hand, it may be instructive to show exactly how some of these alternative proposals violate my axiom scheme: apart from casting light on the motivation for the axioms, this may show how what appear to be coherent and even plausible strategies come apart on close inspection.

The proposed counter-examples, as will become apparent, break into four categories. There are the "wrong-probability" rules, which also require an agent to maximise expected utility but with respect to some probability measure other than the Born rule. There are the 'no-probability" rules, which (purportedly) cannot be represented in terms of expected utilities at all. There are what might be called the "I-don't-want-to-play" rules, which are not so much positive strategies as arguments against the existence of any strategy. And one special group, the contextual strategies, deserve a category of their own.

Branch counting

Description: each branch is given an equal probability, so that if there are N branches following a particular experiment, each branch is given probability $1/N$. Utility is then maximised with respect to this probability.

Origin: Has been reinvented innumerable times, but the first proponent may have been Graham, in DeWitt and Graham (1973).

Rationale: Each branch contains a copy of me; none of them can detect, nor care about, their quantum-mechanical weight; so I should not care about that weight either, and so I have no reason to prefer one over another.

Why it is irrational: The first thing to note about branch counting is that it can't actually be motivated or even defined given the structure of quantum mechanics. There is no such thing as "branch count": as I noted earlier, the branching structure emergent from unitary quantum mechanics does not provide us with a well-defined notion of how many branches there are. All quantum mechanics really allows us to say is that there are *some* versions of me for each outcome.

But within the stylised context of my decision theory, the branch count is defined, so of course (given the representation theorem) the branch counting rule must violate some of my axioms. In fact, it violates the combination of branching indifference and diachronic consistency. For consider two acts $A1$ and $A2$: $A1$ consists of a two-outcome measurement (a spin measurement, say) followed by a reward of utility r in the spin-up branch. $A2$ consists of $A1$, followed by another two-outcome measurement in the spin-up branch. By branching indifference, the agent who gets the reward is indifferent about whether or not he makes a further measurement; by diachronic consistency, then, the original agent is indifferent between $A1$ and $A2$. But the utility of $A1$ (in which there are 2 branches, one of which provides a reward) is $r/2$; the utility of $A2$ is $2r/3$.

The fatness rule

Description: each branch is given a probability proportional to its quantum-mechanical weight multiplied by the mass of the agent in kilograms (such that the total probability is equal to one). Utility is maximised with respect to this probability.

Origin: David Albert (in conversation, and in his contribution to this volume).

Rationale: Albert says, tongue-in-cheek, that an agent should care about branches where he is fatter because "there is more of him" on that branch. He isn't serious, though: the rule is purely presented as a counter-example.

Why it is irrational: It violates diachronic consistency. Albert's agent is (ex hypothesi) indifferent to dieting. But he is not indifferent to whether his future selves diet: he wants the ones on branches with good outcomes to gain weight, and the ones on branches with bad outcomes to lose weight.

This is perhaps a good point to recall the rationale for diachronic consistency: rational action takes place over time and is incompatible with widespread conflict between stages of an agent's life. In the case of the fatness rule, agents have motivation to coerce their future selves — by hiring "minders", say — into dietary programs that they will resist. Multiply this conflict indefinitely many times (for branching is ubiquitous) and rational action becomes impossible.

(To object "maybe rational action is impossible in the Everett interpretation" would, as noted before, be facile. It's perfectly possible for an agent following the Born rule.)

The fake-state rule

Description: The agent maximises expected utilities as for the Born rule, but using a quantum state other than the physically real one.

Origin: Suggested many times in conversation.

Rationale: None in particular, though it is often intended to undermine the connection between the "real" state and the physics.

Why it is irrational: It violates state supervenience. There will be cases where two acts produce the same physical state but where one produces a different fake state than the other. (This is inevitable: any two distinct quantum states are invariant under different sets of transformations.) The fake-state rule will then give the acts different utilities; state supervenience rules this out. Or, put another way: the fake state rule assigns different values to the same physical state under two different descriptions.

Note that it is crucial here — as elsewhere in decision theory — that the agent has a choice between different actions, and therefore between different sets of histories and weights. Of course, in a deterministic universe it is fixed which action will actually occur, but this does not remove the necessity of defining preferences, and hence indirectly probabilities, over a wide range of actions.

The distributive-justice rule

Description: The agent does not maximise expected utilities at all. He treats his various successors in rather the way that a just ruler would treat his various subjects: in particular, he will not allow the suffering of one even if it brings great advantage to others.

Origin: Huw Price (this volume).

Rationale: Any action we choose generates a multitude of individuals; we have a duty to treat them all ethically, and in particular we would not be morally justified in letting one suffer unduly for the others' benefit.

Why it is irrational: The rule is very underspecified, so it isn't easy to answer this, but on natural precisifications it either violates continuity or is not actually a counterexample to the Born rule.

To expand: a large part of what Price wants can be achieved by an appropriate utility function. An agent moved by Price's concerns can drastically increase the disutility of bad consequences and scale down the utility of good consequences, with the effect that trade-offs of the sort he considers get a much lower utility and so will tend to be rejected in favour of more equitable options. There is nothing in Everettian decision theory that prevents an agent from making such modifications to their utility function on recognising the ethical consequences of the Everett interpretation.[13]

If Price wants to hold that *no* amount of suffering, however low-weight the branch on which it occurs, is acceptable, then this strategy will not work, but there is a clash with Continuity. Suppose there are three rewards r_1 and r_2 with $r_1 \succ r_2$, and a (dire) punishment p. Price will prefer r_1 to r_2 but will prefer r_2 to $(1-w)r_1 + wp$, whatever the value of w; clearly this violates continuity.

Now, I think the physical arguments for continuity are pretty unassailable, but it is worth noting that the principle is only really used in my proof precisely to rule out infinite or infinitesimal utilities. (The only other use is for the mathematically convenient but physically tangential purpose of extending the Born rule to the case of infinitely many rewards.) If such utilities are allowed, there is no problem with extending the Born rule to cover even Price's case (though the utility function will have to be modelled in non-standard analysis and the maths will start getting fiddly.) And in fact, precisely the same situation has arisen in *classical* decision theory, and the structure axioms of classical decision theory are selected precisely to rule out the case of infinite (dis)utility.

The variety rule

Description: An agent prefers A to B, but prefers receiving A in half the branches and B in the other half to either A or B.

Author: Suggested in a seminar by Adam Elga in 2004; has not appeared in print as far as I am aware.

Rationale: An agent may regret having to make one choice or another, and may rather like the idea that one version of himself makes one choice, one another. (In Elga's example, a student prefers physics to history but likes

[13] Personally, though, I don't feel inclined to. Call me callous.

both; that student might prefer to do history in one branch, physics in the other.)

Why it is irrational: It either violates diachronic consistency, or it isn't a counter-example to the Born rule.

To expand: suppose you are the agent who chose history. What prevents you changing your mind and switching to physics? It doesn't, after all, hurt your counterpart in the physics branch. This would clearly violate diachronic consistency.

But perhaps you wouldn't choose to switch back. That's to say that although you prefer doing physics to doing history, you prefer doing history *as a result of a situation in which a certain process chose history for you* rather than doing physics *against the result of that process.* In that case, the utility you are assigning to (history-after-process) is higher than the utility you assign to (physics-against-process), and indeed higher than (physics-without-process). The different situations in which you end up doing history count as different rewards.

Exactly analogous situations can arise in classical decision theory. A student might decide that on balance he'd rather do physics than history, but nonetheless resolves to decide by the toss of a coin (because, say, he finds it comforting to have the decision taken from his hands; the reader can probably supply other motivations). That student, again, will place a higher utility on (history after coin toss) than on (physics).

Of course, if every outcome's utility depended sensitively on the circumstances in which that reward arose, decision theory couldn't get off the ground: there would be no way to define probability without being able to have the same reward available in different acts. But again, this is not specific to quantum decision theory.

The anything-goes rule

Description: Not so much a "rule" as a rejection of the need to have one: according to this position, any transitive preference ordering over acts is rationally acceptable.

Origin: Suggested by Tim Maudlin in seminars on multiple occasions; frequently suggested in conversations.

Rationale: Everettian quantum theory is deterministic, and we already have a perfectly acceptable deterministic decision theory: its only axiom is transitivity. So any transitive ordering should be fine.

Why it is irrational: Even in deterministic decision theory, transitivity is not the only constraint. Rational agency is not possible without diachronic consistency; in addition, preference orders have to be defined on actual physical acts, so mathematical modelling of those orders should require an agent to be indifferent between the same state of affairs differently defined. Furthermore, the only interesting decision-theoretic strategies are those which are physically performable in at least an idealised sense. All of the rationality axioms of this paper fit into one of these categories; even in deterministic decision theory, then, they are rationally required.

The curl-up-and-die rule

Description: The converse of the anything-goes rule, this is not so much a "rule" for rational action as the claim that *no* rational strategy is possible in Everettian quantum theory.

Origin: Frequently suggested in conversation.

Rationale: Various; see below.

Why it is irrational: Unless there is something concretely wrong with the Born rule, there is no case to be made that no rational strategy is available: the Born rule is available.

I am aware of two general objections to the rationality of the Born rule, though. The first is that it is rationally compulsory for an agent to weight each branch equally; since the Born rule violates this requirement, it cannot be rational (and if only the Born rule is rational, rationality is impossible in an Everettian universe). Arguments are seldom given for the suggestion that this is a rational requirement (I can see that at best it might be a rational *desideratum*, but it's not at all clear to me why, in a universe where it isn't physically possible to obey the requirement, we should be unable to settle for some second-best option). In any case, though (at the risk of repetitiveness) there is no coherent notion of branch count available in quantum mechanics, so it's not even meaningful to talk of "weighting each branch equally".

The other objection (frequently made in discussions, and made in print by Hemmo and Pitowsky (2007)) is that no strategy can be rational if it can be known in advance by those adopting it that some of them (or some of their successors) will make wrong decisions. So in particular, it is a corollary of the Born rule that an agent measuring a long succession of identical quantum systems should regard the observed frequencies as a guide to what state each system is in; but since all sequences of results occur somewhere, some of the agent's successors will get the wrong outcome.

Now, it is true that some agents will indeed be misled in this way. But there is nothing particularly quantum-mechanical about this. If the universe is spatially infinite (as current observations support), we can guarantee that somewhere in the universe are people as similar to us as you like but whose observed statistics have systematically misled them. Even on Earth, one can fairly easily come up with similar examples. Suppose that the British government declared that it puts some people under (non-covert) surveillance at random, but that there are very few such people: only one in ten million. And suppose it is claimed that the government is lying, and actually puts many more people than that (tens of thousands, say) under surveillance. Then each person in Britain is rational to adopt the strategy: if I am under surveillance, the government is (almost certainly) lying — even though they know that if the government is not lying, five or six people in Britain will be misled into thinking it was.

Ultimately, some people get unlucky. There is no contradiction between this and the rationality of a decision-theoretic strategy, provided that strategy tells us not to care about the unlucky cases. The Born rule tells us exactly that.

Non-contextual rules

Description: An agent's preferences conform to a probability rule that violates the principle of non-contextuality: that is, it assigns different probabilities to the outcomes of a measurement of operator $\widehat{X}$ according to whether or not a compatible operator $\widehat{Y}$ is measured at the same time.

Origin: Various, but a particularly forceful advocacy can be found in (Hemmo and Pitowsky 2007).

Rationale: As is well known, any non-contextual quantum probability rule (and hence, any strategy for rational action expressible in terms of such a rule) can be proved to be the Born rule applied to some (possibly mixed) state.[14] The suspicion, then, is that the decision-theoretic arguments are just a combination of Gleason's theorem (or a relative of it) with an unjustified assumption of non-contextuality.

Why it is irrational: Probably the easiest way to explain what is wrong with non-contextual rules is that they violate State Supervenience. If we regard measurements as physical processes rather than as primitive, which operator(s) are being measured in a given process is dependent on the interests of the experimenter, and cannot simply be read off from the physics. (Consider the Stern-Gerlach experiment, for instance: is it a measurement of spin, or of position?) For a decision rule to be non-contextual, then, is for a rational agent to prefer a given act to the same act (knowably the same act, in fact) under a different description, which obviously violates State Supervenience (and, I hope, is obviously irrational).

[14] This is usually explained in terms of Gleason's Theorem, but this is a rather outdated approach now that POVMs, not PVMs, are widely — and in my view correctly — seen as the best way to represent measurements in quantum theory. Most of the mathematical complexity of Gleason's theorem can be dispensed with if we require our probability function to be defined on POVMs and not just PVMs. See (Caves, Fuchs, Manne, and Renes 2004) for further discussion.

It is fair to note, though, that just as a non-primitive approach to measurement allows one and the same physical process to count as multiple abstractly construed measurements, it also allows one and the same abstractly construed measurement to be performed by multiple physical processes. It is then a non-trivial fact, and in a sense a physical analogue of non-contextuality, that rational agents are indifferent to which particular process realises a given measurement.

In earlier work (Wallace 2003; Wallace 2007) I called this fact *measurement neutrality.* It is indeed a tacit premise in Deutsch's original Deutsch (1999) proof of the Born rule, as I argued in Wallace (2003). In this paper, it is a theorem (a trivial corollary of the main representation theorem, in fact) that measurement neutrality is rationally required. The short answer as to why is that two acts which correspond to the same abstractly construed measurement can be transformed into the same act via processes to which rational agents are indifference. To see the long answer, re-read sections 4–8.

Incidentally, Gleason's theorem (or more accurately its POVM generalisation) is much more directly needed if we wish to generalise the results of this paper to situations where the quantum state is unknown to the agent. The details are somewhat involved; see Wallace (2010) for an account.

10 Conclusion

A rational agent, believing that the Everett interpretation is true and that the quantum state of a given system is $|\psi\rangle$, knows that measurements on that state will generally split his part of the multiverse into multiple branches, with different measurement outcomes, and different versions of the agent, on different branches; he also knows that the relative weights of these branches are given by the Born rule, applied to the post-measurement state of the system and measurement device. Rationality considerations not different in kind to those which apply in single-universe decision making then compel the agent to act as if a set of branches of relative weight w has probability w. In other words, he is rationally required to act as if the Born rule is true.

As I noted in the introduction, my focus here is deliberately narrow and I leave it to other chapters in this volume (and to my own work elsewhere) to make the case that such a result suffices to justify the general role of probability in the Everett interpretation. Yet even on its own terms it is a rather remarkable result, as Deutsch's opening quotation notes, and one which to the best of my knowledge has no analogue outside the branching-universe context.

And how does this result actually come about? The decision-theoretic language in which this paper is written is no doubt necessary to make a properly rigorous case and to respond to those who doubt the very coherence of Everettian probability, but in a way the central core of the argument is not decision-theoretic at all. What is really going on is that the quantum state has certain symmetries and the probabilities are being constrained by those symmetries.

This is actually a throwback to an older idea of probability. Quantitative probability has been concerned with symmetry ever since it was applied to the throw of dice in the 17th century: what makes it reasonable to regard each side of a die as equiprobable is that we have no reason to regard one as more probable than another, and what prevents us having reason is the rotational symmetry of the die that maps one side to another. But real dice — real classical dice, at any rate — must break the symmetry by their initial conditions, or else how in a deterministic universe could the die land one way rather than another. We then have to impose a certain probability distribution on the die's initial conditions, and any prospect of a reductive analysis of probability is lost. In Everettian quantum mechanics, there is no one actual outcome, no requirement for the symmetry to be broken by the actual state of the system, and so a program of deriving the probabilities from the symmetries remains viable. The language of decision theory makes rigorous sense of what such a derivation would look like, and shows — I claim — that the program can indeed be carried out.

Acknowledgments

This work has drawn heavily on conversations and correspondences over a number of years with Harvey Brown, Jeremy Butterfield, David Deutsch, Hilary Greaves, Chris Timpson and Wayne Myrvold, and above all, Simon Saunders.

References

Barnum, H., C. M. Caves, J. Finkelstein, C. A. Fuchs, and R. Schack (2000). Quantum Probability from Decision Theory? *Proceedings of the Royal Society of London A456*, 1175–1182. Available online at http://arXiv.org/abs/quant-ph/9907024.

Caves, C. M., C. A. Fuchs, K. Manne, and J. M. Renes (2004). Gleason-type derivations of the quantum probability rule for generalized measurements. *Foundations of Physics 34*, 193.

Davidson, D. (1973). Radical interpretation. *Dialectica 27*, 313–328.

Davidson, D. (2004). Paradoxes of irrationality. In *Problems of Rationality.* Oxford: Oxford University Press.

Dennett, D. C. (1984). *Elbow Room: the Varieties of Free Will Worth Wanting.* Oxford: Oxford University Press.

Dennett, D. C. (1987). *The intentional stance.* Cambridge, Mass.: MIT Press.

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.

DeWitt, B. and N. Graham (Eds.) (1973). *The many-worlds interpretation of quantum mechanics*. Princeton: Princeton University Press.

Hemmo, M. and I. Pitowsky (2007). Quantum probability and many worlds. *Studies in the History and Philosophy of Modern Physics 38*, 333–350.

Lewis, D. (1974). Radical interpretation. *Synthese 23*, 331–44. Reprinted in David Lewis, *Philosophical Papers*, Volume I (Oxford University Press, Oxford, 1983).

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, P. J. (2005). Probability in Everettian quantum mechanics. Available online at http://phil-sci.pitt.edu.

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

Saunders, S. and D. Wallace (2008). Branching and uncertainty. *British Journal for the Philosophy of Science 59*, 293–305.

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

Wallace, D. (2001). Implications of Quantum Theory in the Foundations of Statistical Mechanics. Available online from http://philsci-archive.pitt.edu.

Wallace, D. (2003). 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. (2005). Language use in a branching universe. Forthcoming; Available online from http://philsci-archive.pitt.edu.

Wallace, D. (2006). Epistemology quantized: circumstances in which we should come to believe in the Everett interpretation. Forthcoming in *British Journal for the Philosophy of Science*. Available online from http://philsci-archive.pitt.edu.

Wallace, D. (2007). Quantum probability from subjective likelihood: Improving on Deutsch's proof of the probability rule. *Studies in the History and Philosophy of Modern Physics 38*, 311–332.

Wallace, D. (2010). *The Everett Interpretation*. Oxford University Press.

Zurek, W. H. and J. P. Paz (1994). Decoherence, chaos and the second law. *Physical Review Letters 72*(16), 2508–2511.